

Probabilistic Generative Models for Inference on Complex Systems

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Martina Contisciani, M. Sc.
aus Fermo, Italien

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 30.01.2024

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Dr. Caterina De Bacco
2. Berichterstatter:	Prof. Dr. Philipp Hennig
3. Berichterstatter:	Prof. Dr. Leto Peel

Disclaimer: this thesis uses Felix Dangel's doctoral thesis template, which is based on Federico Marotta's kaobook class. Furthermore, I would like to thank Kibidi Neocosmos, Laura Iacovissi, Nicolò Zottino, Diego Baptista Theuerkauf, Hadiseh Safdari, Nicolò Ruggeri, and Nathanael Bosch for their help to improve the text.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Caterina De Bacco, for her constant support. Without her patient guidance and honest advice, I would have been completely lost. I am also incredibly thankful for the knowledge and ideas she has shared with me, the time she invested in helping me shape the direction of my research, and the confidence she has instilled in me during the most challenging moments of my PhD journey.

I would also like to thank Philipp Hennig for his steady encouragement and support throughout my doctoral studies, especially when it came to important decisions about my future. I am truly grateful for his presence on both my thesis advisory and examination committees. My gratitude further extends to Bedartha Goswami for his role on my thesis advisory committee, to Anna Levina and Ulrike von Luxburg for serving on my examination committee, and to Leto Peel for reviewing this thesis. I sincerely appreciate their involvement and contribution to my academic journey.

Moreover, I would like to thank the University of Padua for giving me the opportunity to work on my master's thesis in Tübingen. This experience inspired me to pursue a PhD at the Max Planck Institute for Intelligent Systems, which I thank for fostering a stimulating and inclusive environment. I would also like to acknowledge the University of Tübingen, the International Max Planck Research School for Intelligent Systems, and the financial support I received from Cyber Valley.

I am also extremely grateful to all the wonderful people I have had the privilege of working with. This thesis wouldn't have been possible without their collaboration. The meaningful discussions and honest feedback have been a significant part of my personal and professional growth. Additionally, I am especially thankful to my colleagues. The PIO group has been a safe space where I have been able to explore my research ideas, always receiving support and encouragement. I feel fortunate to have pursued my PhD in such a welcoming, healthy, and inspiring environment. I truly believe that I couldn't have come this far without their help and friendship.

While in Tübingen, I also had the chance to make many new friends who added a special touch to the past years, creating unforgettable memories. These people have been an important part of this journey, and I deeply appreciate their friendship, which has brought lots of happiness and comfort, even during tough times. A big thanks also goes to my flatmates. They have been like a second family to me, always there to support me, tolerate my mood swings, and make me feel truly at home.

I am also immensely thankful for the support of my lifelong friends who, despite the physical distance, consistently made an effort to stand by me and even visit. Many precious moments together had to be sacrificed for this achievement, and I am grateful for their patience and understanding.

Lastly, I would like to extend my gratitude to my family, who always encouraged me to follow my ambitions. Their love never fails to make me feel special. Also, I wouldn't thank Nico enough for his emotional support throughout this journey. I am thankful to him for believing in me even more than I do myself, for always being there despite the distance, for listening, giving valuable advice, and making me laugh.

Martina Contisciani
Tübingen, February 9, 2024

Abstract

Network models are powerful and flexible tools to represent the complex interactions between individual elements in diverse domains. They offer scientists and practitioners willing to exploit the growing abundance of networked datasets meaningful insights into the fundamental patterns underlying such interactions. A popular approach to identify these hidden structures is that of generative models, in particular latent variable models: probabilistic models that introduce latent variables to incorporate domain knowledge, capture complex interactions, and uncover statistically meaningful network structures. Existing methods are frequently insufficient to capture the complexity of real-world data, and they often do not provide a general framework to fully leverage the additional information carried within the data, such as edges and nodes metadata or higher-order interactions.

In this thesis, we present principled and efficient approaches that aim to broaden the range of techniques available for modelling complex networks. Specifically, we work in three principal directions: i) developing flexible methods to perform inference on attributed multilayer networks, ii) exploring innovative theoretical perspectives for incorporating reciprocity and loosening the assumption of conditional independence in network models, iii) designing foundational models to characterize the structural organization of higher-order data.

We first extend standard generative models for the analysis of multilayer networks to integrate node metadata into the inference process with the network topology. In addition to applying these methods to already explored real-world data, such as social and biological networks, we introduce this methodology to another field for the first time, that is patent citation networks. We show how incorporating additional information not only boosts performance, but also leads to more interpretable structures.

Next, we propose approaches to handle the pairwise dependencies between two directed edges connecting node pairs, which come with the relaxation of the assumption that edges are independent of each other. We demonstrate the flexibility and relevancy of our mathematical frameworks in various contexts, such as the analysis of dynamic networks, identification of anomalies, and estimation of unobserved network structures using multiple reports. By explicitly accounting for reciprocity, it improves edge prediction and network reconstruction, while also shedding light on the underlying mechanisms driving edge formation.

Finally, we present principled methods to define and identify the mesoscale organization of higher-order data. We evaluate their effectiveness on a variety of small- and large-scale real-world systems. Notably, these models display good performance in effectively retrieving both robust and flexible community structures, while reliably predicting higher-order interactions of arbitrary size. As an additional contribution, we present a newly developed Python library specifically designed for analyzing data with higher-order interactions.

This work thus introduces cutting-edge techniques that go beyond what has been previously established in the field of network inference and contribute to the enhancement of the current literature. These developed approaches account for the additional complexity present in real-world systems, enabling a more profound understanding of data across a range of different disciplines.

Zusammenfassung

Netzwerkmodelle sind eine leistungsfähige und flexible Methode zur Darstellung der komplexen Interaktionen in verschiedenen Anwendungsbereichen. Sie bieten Wissenschaftlern und Anwendern, die bereit sind die wachsende Fülle vernetzter Datensätze zu nutzen, aussagekräftige Einblicke in die zugrundeliegenden Muster solcher Interaktionen. Ein beliebter Ansatz zur Ermittlung dieser verborgenen Strukturen sind generative, probabilistische Modelle in welchen latenten Variablen eingeführt werden, um Domänenwissen einzubeziehen, komplexe Interaktionen zu erfassen und statistisch aussagekräftige Netzwerkstrukturen aufzudecken. Bestehende Methoden sind häufig unzureichend, um die Komplexität realer Daten zu erfassen, und sie bieten oft keinen umfassenden Rahmen, um die zusätzlichen Informationen, welche in den Daten enthalten sind, wie z.B. Metadaten oder Interaktionen höherer Ordnung, vollständig genutzt werden können.

In dieser Arbeit stellen wir fundierte und effiziente Ansätze vor, welche darauf abzielen, die Palette der verfügbaren Methoden zur Modellierung komplexer Netzwerke zu erweitern. Konkret arbeiten wir in drei Hauptrichtungen: i) die Entwicklung flexibler Methoden zur Durchführung von Inferenzen auf attribuierten mehrschichtigen Netzwerken, ii) die Erforschung innovativer theoretischer Perspektiven zur Miteinbeziehung von Reziprozität und zur Lockerung der Annahme bedingter Unabhängigkeit in Netzwerkmodellen, und iii) der Entwurf grundlegender Modelle zur Charakterisierung struktureller Organisation von Daten höherer Ordnung.

Wir erweitern zunächst bekannte generative Analysemodelle für mehrschichtige Netzwerke, um Knoten-Metadaten in den Inferenzprozess mit der Topologie zu integrieren. Wir wenden diese Methoden auf verschiedenen realen Daten an, sowohl auf bereits erforschten Netzwerken als auch auf zum ersten Mal auf Patentzitiernetzwerken. Wir zeigen auf, wie zusätzliche Informationen nicht nur Vorhersagen verbessern, sondern auch zu besser interpretierbaren Strukturen führen.

Anschließend schlagen wir Ansätze zur Handhabung der paarweisen Abhängigkeiten zwischen zwei gerichteten, Knotenpaare verbindenden, Kanten vor, welche mit der Lockerung der Unabhängigkeitsannahme zwischen Kanten einhergeht. Wir zeigen die Flexibilität und Relevanz unserer Methodik in verschiedenen Kontexten auf, z.B. bei der Analyse dynamischer Netzwerke, bei der Identifizierung von Anomalien und bei der Schätzung unbeobachteter Netzwerkstrukturen unter Verwendung mehrerer Berichte. Durch die explizite Berücksichtigung der Reziprozität wird die Vorhersage von Kanten und die Rekonstruktion von Netzwerken verbessert, während gleichzeitig die zugrunde liegenden Mechanismen der Kantenbildung verdeutlicht werden.

Schließlich stellen wir fundierte Methoden zur Identifizierung der mesoskaligen Organisation von Daten höherer Ordnung vor. Wir werten ihre Effektivität auf einer Vielzahl von kleinen und großen realen Systemen aus. Diese Modelle zeigen eine gute Leistung bei der Suche nach robusten und flexiblen Gemeinschaftsstrukturen sowie bei der zuverlässigen Vorhersage von Interaktionen beliebiger Größe auf. Zusätzlich stellen wir eine neu entwickelte Python-Bibliothek vor, welche speziell für die Analyse von Daten mit Interaktionen höherer Ordnung entwickelt wurde.

Diese Arbeit führt somit innovative Techniken ein, die über das hinausgehen, was bisher auf dem Gebiet der Netzwerkinferenz entwickelt wurde, und trägt zur Weiterentwicklung der aktuellen Forschungsliteratur bei. Alles in allem ermöglichen die entwickelten Ansätze ein tieferes Verständnis von realen Daten in verschiedenen Anwendungsbereichen.

Table of Contents

Acknowledgments	v
Abstract	vii
Zusammenfassung (German Abstract)	ix
Table of Contents	xi
1 Introduction	1
1.1 Overview & Motivation	1
1.2 Outline	2
2 Background	5
2.1 Network data	5
2.1.1 Single-layer networks	5
2.1.2 Multilayer networks	6
2.1.3 Higher-order networks	7
2.2 Statistical models and inference	8
2.2.1 Probabilistic generative models	9
2.2.2 Expectation-Maximization	11
2.2.3 Variational Inference	12
2.3 Community detection models	14
2.3.1 Community structures	15
2.3.2 Stochastic Block Model	16
2.3.3 Multitensor	18
3 Published Work	19
3.1 Inference on attributed multilayer networks	19
3.1.1 Community detection with node attributes in multilayer networks	20
3.1.2 Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships	20
3.2 Reciprocity and the relaxation of the conditional independence assumption	21
3.2.1 Generative model for reciprocity and community detection in networks	22
3.2.2 Community detection and reciprocity in networks by jointly modelling pairs of edges	23
3.2.3 Reciprocity, community detection, and link prediction in dynamic networks	23
3.2.4 Anomaly, reciprocity, and community detection in networks	24
3.2.5 Latent network models to account for noisy, multiply reported social network data	25
3.3 Community detection and the analysis of higher-order data	26
3.3.1 Inference of hyperedges and overlapping communities in hypergraphs	26
3.3.2 Community detection in large hypergraphs	27
3.3.3 Hypergraphx: a library for higher-order network analysis	27

4 Discussion & Conclusion	29
4.1 Inference on attributed multilayer networks	29
4.2 Reciprocity and the relaxation of the conditional independence assumption	31
4.3 Community detection and the analysis of higher-order data	34
4.4 Conclusion	35
Bibliography	37
A Appendix	47
A.1 Published papers	47

1 Introduction

1.1 Overview & Motivation

Over the past few decades, the study of networks has gained prominence across various disciplines as a valuable method for analyzing complex relational data. The origins of this approach can be traced back to 1736 when Euler utilized the mathematical representation of vertices and edges to successfully solve the famous Seven Bridges of Königsberg problem [55]. Since then, this formalism has found widespread use in describing real-world data from diverse domains. For instance, social networks have proven to be a fruitful tool for representing and understanding social relationships like friendships and collaborations, along with their implications [26, 114, 151]. Similarly, biological networks offer valuable insights into complex systems such as protein-protein interactions [90, 147], brain networks [27, 50], and ecological food webs [12, 49], providing a deeper understanding of biological processes. Moreover, networks play a crucial role in analyzing the internet [59, 155, 156], transportation systems [10, 72, 94], and power grids [4, 41, 115], enabling researchers to optimize their efficiency and robustness. Networks have also proven invaluable in studying epidemic spread [85, 117, 148], facilitating prediction and control of disease outbreaks by analyzing epidemiological networks and developing contagion spread models.

Despite the different characteristics and generation processes of the networks within each subfield, a crucial finding of network science is the shared architecture among networks arising in diverse domains [11]. This renders networks a versatile and general framework, allowing us to use a common set of mathematical tools to explore systems from diverse domains. Another aspect to consider during the analysis of such systems is that data collection methods are advancing due to the technological progress, leading to the acquisition of more comprehensive data. For instance, many relational datasets now come with additional information attached to nodes and edges [87, 88], offering the opportunity to incorporate this metadata in the analysis of such systems. Additionally, recent observations have revealed that many real-world interactions are not independent of each other; instead, they encompass dependencies like reciprocity [63], or they occur in higher-order forms rather than just pairwise connections [13]. Consequently, models and representations employed for analyzing this data must evolve to accurately capture this complexity, enabling more profound understanding of real-world systems. As such, the objective of this thesis is to introduce a range of alternative statistical methodologies specifically designed to provide substantial insights into comprehending complex networks as they exist in reality.

Among the numerous research directions in network analysis, this thesis focuses on network inference, the process of learning the properties of complex networks from data [58]. To achieve this, we employ probabilistic generative models [69], which are statistical tools that describe how data may be generated through underlying variables and parameters. These methods utilize probability distributions to model the interactions, allowing us to capture the inherent uncertainty and complexity that often characterize real-world networks. In particular, we develop latent variable models [56], which are flexible and powerful probabilistic models that incorporate hidden variables that are not directly observed in the data. These latent variables allow the injection of domain knowledge into the theoretical framework and the uncovering of statistically meaningful network structures. For example, a common assumption in network inference is the belief that nodes'

interactions are driven by hidden communities. In this framework, the latent variables represent the nodes' community memberships and the interplay among these communities, with the aim to infer these quantities from the data [9]. This problem is known as community detection [60], and it is one of the foundational elements underpinning the models introduced in this study.

Over the past few decades, there has been a rapid increase in the use of latent variable models within network science for analyzing real-world data. Nevertheless, as networks become more sophisticated, the current techniques fall short in fully capturing the wide range of complexities present in such data. In fact, existing methods tend to oversimplify the dependencies between node interactions, and they often lack the flexibility needed to effectively and simultaneously utilize the extra information contained in the data, such as both node and edge metadata. These limitations highlight the need for new approaches that can comprehensively handle the many complexities of real-world data. In response to these shortcomings, this thesis introduces a set of statistical methods specifically tailored to address the multifaceted challenges posed by real-world networks. More precisely, we tackle three distinct facets of complexity: i) we develop flexible methods which include node and edge metadata to perform inference on networks [38, 74], ii) we propose innovative approaches for incorporating reciprocity in latent network models through the relaxation of the conditional independence assumption [39, 44, 135–137], iii) we design foundational models to characterize the structural organization of higher-order data, i.e., systems which involve interactions between groups of nodes of any arbitrary size [37, 96, 134].

All in all, by employing alternative probabilistic approaches and integrating domain-specific knowledge, the proposed methods strive to expand the existing toolbox of techniques available for modeling complex networks. By pushing the limits of latent variable modeling, this thesis seeks to equip professionals and researchers with advanced instruments capable of capturing the multifaceted nature of modern network data. This effort ultimately contributes to a more sophisticated understanding of complex relationships and hidden structures.

1.2 Outline

The rest of the thesis is structured into three chapters: an introduction to the setting and related concepts, such as data, methodology, and community detection; the scientific contributions, and a discussion of their impact and future directions. As a cumulative thesis, the published papers are collected in [Appendix A](#).

Chapter 2 provides background information useful to understand the context into which the methods outlined in this thesis align, that is, statistical inference of network data. First, **Section 2.1** introduces network data and the diverse mathematical structures utilized for their analysis. This includes an exploration of single-layer networks, multilayer networks, and higher-order networks – each tailored to capture specific interaction patterns. Subsequently, **Section 2.2** delves into the statistical framework employed for the analysis of these networks, and the techniques applied to infer the latent variables. Lastly, **Section 2.3** provides an overview of the community detection problem, along with an explanation of two state-of-the-art generative models addressing this challenge.

Chapter 3 presents the published work, organized into the three main directions explored in this thesis. **Section 3.1** showcases the publications centered on performing inference on attributed

multilayer networks – networks enriched with node and edge metadata. This section includes the following work:

- ▶ **M. Contisciani**, E. A. Power, and C. De Bacco. “Community detection with node attributes in multilayer networks”. *Scientific Reports* 10.1 (2020), page 15736. [38]
- ▶ K. Higham, **M. Contisciani**, and C. De Bacco. “Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships”. *Technological Forecasting and Social Change* 179 (2022), page 121628. [74]

In contrast, [Section 3.2](#) illustrates the contributions addressing pairwise dependencies between directed edges that connect pairs of nodes, extending beyond the concept of conditional independence. This section includes the following publications:

- ▶ H. Safdari*, **M. Contisciani***, and C. De Bacco. “Generative model for reciprocity and community detection in networks”. *Physical Review Research* 3.2 (2021). *Contributed equally. [135]
- ▶ **M. Contisciani**, H. Safdari, and C. De Bacco. “Community detection and reciprocity in networks by jointly modelling pairs of edges”. *Journal of Complex Networks* 10.4 (2022). [39]
- ▶ H. Safdari, **M. Contisciani**, and C. De Bacco. “Reciprocity, community detection, and link prediction in dynamic networks”. *Journal of Physics: Complexity* 3.1 (2022). [136]
- ▶ H. Safdari, **M. Contisciani**, and C. De Bacco. “Anomaly, reciprocity, and community detection in networks”. *Physical Review Research* 5.3 (2023). [137]
- ▶ C. De Bacco, **M. Contisciani**, J. Cardoso-Silva, H. Safdari, G. Lima Borges, D. Baptista, T. Sweet, J.-G. Young, J. Koster, C. T. Ross, R. McElreath, D. Redhead, and E. A. Power. “Latent network models to account for noisy, multiply reported social network data”. *Journal of the Royal Statistical Society Series A: Statistics in Society* 186.3 (2023). [44]

Finally, [Section 3.3](#) expounds work focused on the analysis and structural characterization of higher-order networks. This section includes the following papers:

- ▶ **M. Contisciani**, F. Battiston, and C. De Bacco. “Inference of hyperedges and overlapping communities in hypergraphs”. *Nature Communications* 13.1 (2022). [37]
- ▶ N. Ruggeri, **M. Contisciani**, F. Battiston, and C. De Bacco. “Community detection in large hypergraphs”. *Science Advances* 9.28 (2023). [134]
- ▶ Q. F. Lotito, **M. Contisciani**, C. De Bacco, L. Di Gaetano, L. Gallo, A. Montresor, F. Musciotto, N. Ruggeri, and F. Battiston. “Hypergraphx: a library for higher-order network analysis”. *Journal of Complex Networks* 11.3 (2023). [96]

[Chapter 4](#) summarizes the findings of each scientific contribution and combines them together to provide a broader discussion of the different research directions investigated. This last chapter also highlights the impact and the implications of the achieved results in relation to the latest developments in the field. Moreover, the conclusions delineate potential avenues for future research, both within the specific areas under investigation and in the broader scope of developing latent variable models for network analysis.

2 Background

In this chapter, we present an overview of the key elements and the structural framework that underlie the models outlined in this thesis. To begin, [Section 2.1](#) delves into the realm of network data, elucidating various mathematical representations utilized to depict the different complexities of real-world data. Following that, [Section 2.2](#) introduces the statistical methodology employed for the analysis of these networks, that is probabilistic generative models. In particular, we focus on latent variable models, whose main assumption relies on the belief that real-world networks can be explained through a compact set of latent variables, which must be inferred from the data. To illustrate this concept, [Section 2.3](#) expounds community detection models. In this context, the latent variables represent the communities to which nodes belong, and their inference reveals the hidden structures that influence node interactions.

2.1 Network data

Real-world data often exhibit the characteristics of complex systems [146] – systems that are challenging to model due to the intricate interplay of dependencies, competitions, and relationships existing among their various components or between the system itself and its surrounding environment. Notable examples encompass the human brain, infrastructure networks, and communication systems, among others. In the field of network science, such systems are effectively represented using networks, also known as graphs. In this representation, nodes within the network correspond to individual components, while the links between nodes represent their interactions [105]. By transforming real-world data into network data, researchers and practitioners can employ a powerful methodology to understand the underlying structure and behavior of complex systems.

Within this section, we describe the mathematical framework, fix the notation, and present various network representations. We first introduce single-layer networks, which help us understand the key characteristics of networks and the concept of attributed networks. They also serve as the starting point for more complex network models, such as multilayer networks and the emerging notion of higher-order networks.

2.1.1 Single-layer networks

A *single-layer network*, often referred to as *network* or *graph*, is a collection of nodes (also called vertices) that are connected by edges (also known as links). We denote a network as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, N\}$ constitutes the set of nodes, and $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\}$ represents the set of edges. Mathematically, a graph is represented through its adjacency matrix $A = \{A_{ij}\} \in \mathbb{R}^{N \times N}$, wherein A_{ij} denotes the edge from node i to node j ($i \rightarrow j$). If $A_{ij} \neq 0$, then i and j are neighbors or adjacent. The total number of neighbors of a node i , essentially the number of its connections, is referred to as its degree. In real-world networks, the overall count $|\mathcal{E}|$ of non-zero entries generally scales linearly with the number of nodes N , making it significantly smaller than the potential maximum connections $N \times N$. This characteristic is known as sparsity and has implications for efficiency, computational complexity, and the interpretation of network structures [11].

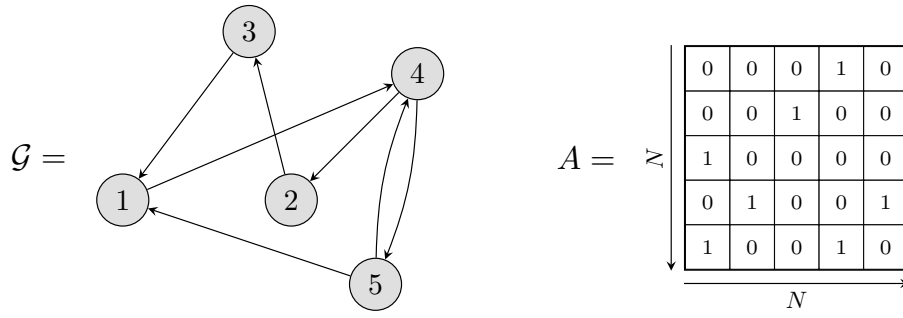


Figure 2.1: Illustration of a directed single-layer network. (left) The graphical representation of the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and (right) its corresponding adjacency matrix A . The node set is given by $\mathcal{V} = \{1, 2, 3, 4, 5\}$, and the edge set is denoted by $\mathcal{E} = \{(1, 4), (2, 3), (3, 1), (4, 2), (4, 5), (5, 1), (5, 4)\}$. This is a binary (or unweighted) network, where the edges indicate either the presence ($A_{ij} = 1$) or the absence ($A_{ij} = 0$) of interactions.

A network is undirected when its edges do not have a direction, indicating a reciprocal relation, which translates to $(i, j) \in \mathcal{E} \leftrightarrow (j, i) \in \mathcal{E}$. Consequently, $A_{ij} = A_{ji}$, leading to a symmetric adjacency matrix. Conversely, directed networks encode directionality to their links: $i \rightarrow j$ implies that $(i, j) \in \mathcal{E}$, which can be different from $j \rightarrow i$. In our contributions, we mainly focus on the broader context of directed networks, and we assume that a node cannot establish a link with itself, thereby excluding self-loops. A simplified representation of this kind of network can be observed in Figure 2.1. In this example, we show a binary (or unweighted) network, where the edges indicate either the presence ($A_{ij} = 1$) or the absence ($A_{ij} = 0$) of interactions. Nonetheless, real-world interactions frequently incorporate varying weights, such as representing the number of calls between individuals or the electric current flowing through a transmission line within a power grid. Such networks are referred to as weighted networks. The majority of our work focuses on developing models capable of handling nonnegative discrete weights, thus taking in input $A = \{A_{ij}\} \in \mathbb{N}_0^{N \times N}$.

Real-world datasets often encompass also other types of information. Within this thesis, we define an *attributed network* as one in which nodes are associated with supplementary information or attributes (also referred to as covariates). We represent this information using a design matrix $X \in \mathbb{R}^{N \times P}$, in which each row corresponds to an individual node, and the columns denote attributes alongside their specific values for that node. The incorporation of such metadata into networks adds another layer of information, that can be crucial for understanding the network's structure, dynamics, and underlying properties.

2.1.2 Multilayer networks

In most real-world systems, a set of entities interact with each other in complicated patterns that can encompass multiple types of relationships, change in time, and include other types of complexities [88]. Single-layer networks fall short in grasping this multifaceted complexity as they treat all interactions uniformly, neglecting supplementary details about the nature of these interactions. In contrast, *multilayer networks* explicitly encompass diverse types of interactions and provide a suitable framework for describing systems linked through various forms of connections. In this context, each interaction is represented as a distinct layer and a given node can engage in various types of interactions, resulting in different sets of neighbors within each layer [24]. For

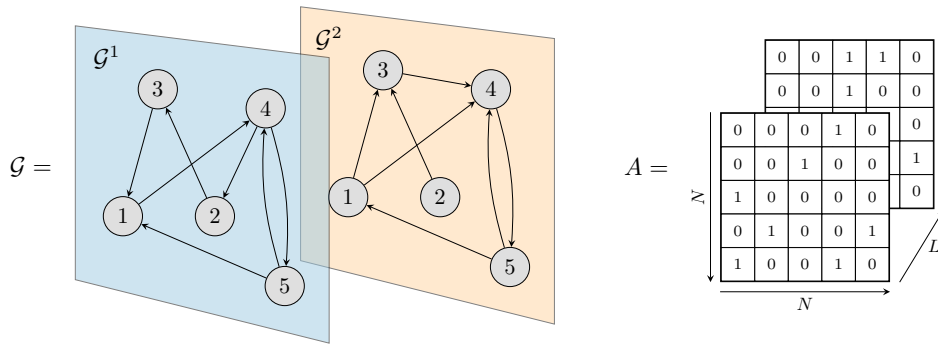


Figure 2.2: Illustration of a directed multilayer network. (left) The graphical representation of the multilayer network $\mathcal{G} = \{\mathcal{G}^\ell(\mathcal{V}, \mathcal{E}^\ell)\}_{1 \leq \ell \leq 2}$, and (right) its corresponding adjacency tensor A . The node set is given by $\mathcal{V} = \{1, 2, 3, 4, 5\}$, and the edge sets are denoted by $\mathcal{E}^1 = \{(1, 4), (2, 3), (3, 1), (4, 2), (4, 5), (5, 1), (5, 4)\}$ and $\mathcal{E}^2 = \{(1, 3), (1, 4), (2, 3), (3, 4), (4, 5), (5, 1), (5, 4)\}$ for the two layers respectively. This is a binary (or unweighted) multilayer network, where the edges indicate either the presence ($A_{ij}^\ell = 1$) or the absence ($A_{ij}^\ell = 0$) of interactions.

instance, in social networks individuals may share diverse types of relationships, such as friendship, family ties, and professional connections. In transportation networks, instead, different layers might correspond to diverse modes of transport, such as roads, railways, and air routes.

From a mathematical perspective, a multilayer network can be expressed as a multilayer graph $\mathcal{G} = \{\mathcal{G}^\ell(\mathcal{V}, \mathcal{E}^\ell)\}_{1 \leq \ell \leq L}$ defined on a set \mathcal{V} of N vertices shared across $L \geq 1$ layers. Specifically, this representation corresponds to a particular case of multilayer network known as multiplex network [14, 144], where nodes are common to all layers, and interactions take place only within layers without spanning across them. In this context, each layer $\ell \in \{1, \dots, L\}$ can be represented as a graph $\mathcal{G}^\ell(\mathcal{V}, \mathcal{E}^\ell)$ with an associated adjacency matrix $A^\ell = \{A_{ij}^\ell\} \in \mathbb{R}^{N \times N}$, wherein A_{ij}^ℓ indicates the strength of the connection of type ℓ from node i to node j . This multiplex system can be entirely described using a 3-dimensional adjacency tensor A , having dimensions $L \times N \times N$. An illustrative example of a 2-layer network is depicted in Figure 2.2. Similarly to single-layer networks, multilayer networks can be directed or undirected, and their edges may be binary or weighted. Furthermore, an *attributed multilayer network* denotes a multilayer structure where the nodes carry additional metadata.

Real-world data can also encode interactions that change over time. For instance, in financial markets, relationships between assets or trading entities change as market conditions fluctuate. These types of interactions can be represented by *temporal networks* [79], where the links between nodes change at different time points, reflecting the dynamic nature of relationships. Temporal networks, also known as *dynamic networks*, can be conceptualized as multilayer networks: each layer ℓ corresponds to a snapshot of the network at a specific time step t , and the edges in each layer capture the state of the network at that moment. This representation allows us for a more comprehensive understanding of how interactions evolve and how the network structure adapts to various temporal dynamics.

2.1.3 Higher-order networks

Networks, whether they are single-layer or multilayer, face a significant limitation as they exclusively capture pairwise interactions [13, 15, 18]. Nevertheless, real-world systems from various domains, including social systems [16], biology [138], ecology [71], and neuroscience [126], exhibit interactions

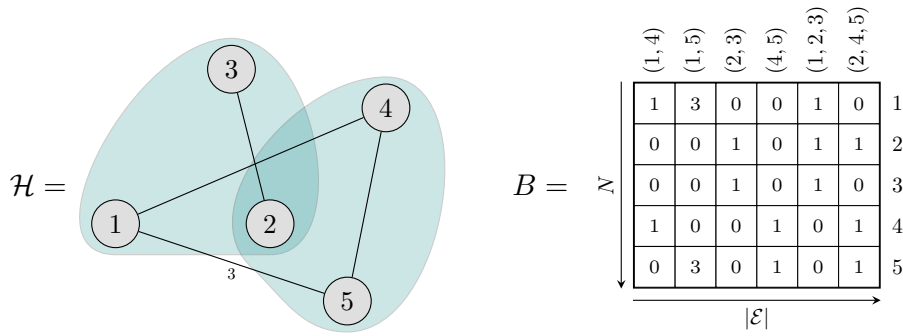


Figure 2.3: Illustration of a higher-order network. (left) The graphical representation of the hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, and (right) its corresponding incidence matrix B . The node set is given by $\mathcal{V} = \{1, 2, 3, 4, 5\}$, and the hyperedge set is denoted by $\mathcal{E} = \{(1, 4), (1, 5), (2, 3), (4, 5), (1, 2, 3), (2, 4, 5)\}$. The 2-dimensional interactions (edges) are represented with a black line, and the 3-dimensional interactions are shown with a colored set. This is a weighted hypergraph where the hyperedge weights are denoted only if different from 1.

involving three or more system units at a time. For instance, in co-authorship networks, scientific papers might involve more than two authors collaborating. Similarly, in protein interaction networks, interactions among proteins manifest as protein complexes, wherein multiple proteins bind together. Consequently, these higher-order systems are most appropriately described using different mathematical frameworks, such as *hypergraphs* [17], capable of representing relationships among any number of nodes through hyperedges of varying dimensions. Embracing the inherent higher-order nature of these systems enhances our modeling capabilities, leading to a more profound understanding of their complex structure.

Formally, a hypergraph extends the concept of a graph and is represented as $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} and \mathcal{E} denote the node set and the hyperedge set, respectively. Each hyperedge $e \in \mathcal{E}$ is a non-empty subset of \mathcal{V} , representing a higher-order interaction between an arbitrary number $|e|$ of nodes. The set of all possible hyperedges among nodes in \mathcal{V} is represented as Ω . We denote by D the maximum hyperedge size, which can be set up to a maximum value of $D = N$. In the context of networks, $D = 2$ and $\Omega = N \times N$. A hypergraph is described using an adjacency tensor $A = \{A_e\} \in \mathbb{R}^\Omega$, where the entry A_e represents the weight of the d -dimensional edge $e \in \Omega$, with $d = |e|$. For example, in the case of co-authorship interactions, $A_{i,j,h}$ might denote the number of papers written together by the authors i, j and h . Another way to characterize a hypergraph is through the incidence matrix $B = \{B_{ie}\} \in \mathbb{R}^{N \times |\mathcal{E}|}$. In this representation, each entry B_{ie} indicates the weight associated with the hyperedge e that includes the node i . Our study focuses on undirected hypergraphs with nonnegative discrete weights, and we provide an illustrative example in Figure 2.3.

2.2 Statistical models and inference

In the realm of network science theory, the key to gain a profound understanding of the intricate structure and behavior of complex systems lies in the analysis of the networks that serve as representations of these systems [11]. A powerful approach for examining and comprehending these networks involves the application of probabilistic generative models. These models provide a statistical framework that not only accommodates the underlying generative process but also effectively handles the inherent uncertainty present in the observed data.

In this section, we explore the framework of probabilistic generative models, with a specific emphasis on latent variable models. These models offer a probabilistic representation of the data in terms of both observable and hidden variables, where the latter capture concealed patterns within the data [20]. Furthermore, we provide detailed explanations for two methods used in this thesis to infer the latent variables: the Expectation-Maximization (EM) algorithm and the Variational Inference (VI) approach. EM is an iterative algorithm used to provide point estimates of the latent variables, while VI serves as a technique for approximating their posterior distribution.

2.2.1 Probabilistic generative models

Probabilistic generative models are a class of statistical models that aim at representing the data in terms of an unknown and approximate probability distribution $P(A)$. In other words, these models seek to learn the mechanisms by which networks are generated, enabling them to produce new data that closely resemble what has been observed. Generative models are powerful methods because they are extensible and can be easily adapted to explicitly encode specific hypotheses and assumptions about complex systems and how we observe them [119].

An example of a probabilistic generative model in network science is the Erdős–Rényi model [52, 53], employed to generate random graphs. This model assumes that graphs with a fixed vertex set \mathcal{V} and a predetermined number of edges $|\mathcal{E}|$ are equally probable, and a graph is chosen uniformly at random from the collection of all graphs containing $|\mathcal{V}|$ nodes and $|\mathcal{E}|$ edges. A related model, introduced by Gilbert [67], generates a random graph by independently selecting each edge with a constant probability p for its existence, regardless of the presence or absence of other edges.

Although these models serve as valuable tools for analysis, their underlying assumptions are somewhat too simplistic to effectively capture the complexities present in real-world data. It is, indeed, an oversimplification to assume that interactions within a real system occur randomly, as they are often influenced by hidden mechanisms, such as group membership or dynamics. To properly address such interdependencies, we turn to *latent variable models*, a category of probabilistic generative models that attempt to explain a complex observed dataset in terms of simpler, but unobserved, patterns [20]. These patterns are explicitly modeled through latent variables, denoted as Θ , and by estimating them, one can gain insights into the underlying network structures. For example, if we suppose that these latent variables represent node memberships, inferring them could unveil, for instance, that two individuals interact more frequently because they belong to the same group of friends.

The development of latent variable models, and probabilistic generative models in general, follows an iterative process, as outlined in Figure 2.4. Initially, we make assumptions about the hidden variables that might influence the data generation process, and we represent these relationships in a joint probability distribution of both observed and hidden random variables [20], denoted as $P(A, \Theta)$. Subsequently, when we have an observed dataset, such as a graph \mathcal{G} (or equivalently its adjacency matrix A), we perform inference on these variables. Once these variables are learned, the model can be used to predict missing data or generate new synthetic data that align with the underlying generative process. This serves as a model validation, as if the synthetic data exhibits properties similar to those of the observed data, it signifies that our assumptions are correct. Conversely, if they differ significantly, it prompts a reevaluation and potential redefinition of the model assumptions.

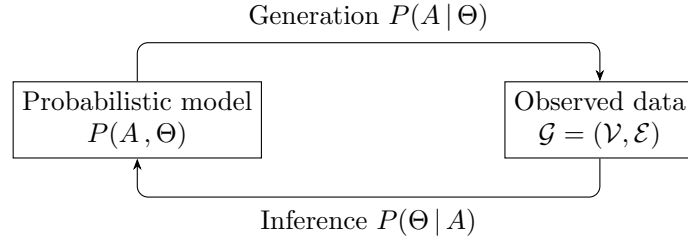


Figure 2.4: Graphical representation of the latent variable models framework. The aim is to posit a probabilistic model that elucidates the generative process underlying the observed data A . This model incorporates latent variables Θ , which capture the assumptions about the hidden mechanisms responsible for shaping the interactions. These latent variables are subsequently inferred using an observed dataset, such as a graph \mathcal{G} or its adjacency matrix A , and then employed to generate new data that align with the underlying generative process.

Based on the underlying generative process assumptions, the joint distribution can be factorized as $P(A, \Theta) = P(A | \Theta) P(\Theta)$. In this formulation, $P(A | \Theta)$ represents the likelihood of the model, while $P(\Theta)$ characterizes the prior distribution of the latent variables. To enhance the manageability of these models, a commonly employed assumption is that of conditional independence: conditioned on the latent variables, the observed variables become independent of each other. In the context of network modeling, where the observed variables are the edges of the graph, this implies that all interactions are independent and identically distributed given the latent variables, resulting in the following expression:

$$P(A, \Theta) = \prod_{i,j} P(A_{ij} | \Theta) P(\Theta). \quad (2.1)$$

This definition is quite broad, and the appropriate choice of $P(A_{ij} | \Theta)$ and $P(\Theta)$ hinges on the available data and the specific research questions being pursued. As an example, when our graph consists of binary edges, $P(A_{ij} | \Theta)$ might take the form of a Bernoulli distribution. Conversely, if the edges are characterized by nonnegative discrete weights, then $P(A_{ij} | \Theta)$ could be modeled using a Poisson distribution. In this thesis, irrespective of the chosen distribution, we assume that the likelihoods are fully parametrized through the latent variables Θ .

Once a probabilistic model has been defined, the next step involves choosing the method for inferring the latent variables Θ . Depending on the specific objectives and the model's computational feasibility, one may be interested in either inferring single-point estimates or estimating the full posterior distribution of Θ . In the former scenario, a suitable approach is to employ a Maximum Likelihood Estimate (MLE), where $\hat{\Theta}$ represents the values that maximize the probability of the observed data under the assumed statistical model:

$$\hat{\Theta}_{MLE} = \arg \max_{\Theta} P(A | \Theta). \quad (2.2)$$

Conversely, an alternative approach is to opt for a Maximum A Posteriori (MAP) estimate, which seeks the most probable values by considering both the likelihood of the data and the prior probability distribution over the latent variables:

$$\begin{aligned} \hat{\Theta}_{MAP} &= \arg \max_{\Theta} P(\Theta | A) \\ &\propto \arg \max_{\Theta} P(A | \Theta) P(\Theta). \end{aligned} \quad (2.3)$$

A method used to solve these maximization problems is the Expectation-Maximization algorithm, that will be introduced in the following subsection. Similarly, the last subsection will delve into Variational Inference, a method employed to estimate full posterior distributions within the context of a Bayesian framework.

2.2.2 Expectation-Maximization

The *Expectation-Maximization (EM)* algorithm is an iterative statistical method designed to compute point estimates for the parameters Θ of probabilistic models that rely on unobserved or missing variables Z [48, 101, 103]. These variables are introduced into the model to simplify the expression of the likelihood $P(A | \Theta)$, which is often complicated and challenging to manipulate. In essence, the EM algorithm alternates between two main steps: first, it estimates the expected values of the unknown variables Z (E step), and then it updates the model parameters to maximize the likelihood of the observed data (M step). In our specific context, the probabilistic models are entirely characterized by the latent variables Θ , which can be considered the only model parameters that we aim to infer.

The likelihood in Equation (2.2) can be reformulated by using the logarithm and explicitly account for the unknown variables Z as follows:

$$\begin{aligned}\mathcal{L}(\Theta) &:= \log P(A | \Theta) \\ &= \log \sum_Z P(A, Z | \Theta).\end{aligned}\tag{2.4}$$

In this representation, we treat Z as discrete random variables, but this concept can be extended to continuous variables by simply substituting the summation with an integral. Explicitly maximizing $\mathcal{L}(\Theta)$ can be challenging due to the presence of the logarithm of a sum. In such a setting, the EM algorithm offers an efficient method to solve this problem by approximating $\mathcal{L}(\Theta)$ with a lower bound that can be maximized to find the estimates for the model parameters. To construct this lower bound, let's introduce a distribution q over the possible values of Z , satisfying $\sum_Z q(Z) = 1$ and $q(Z) \geq 0$. By using the Jensen's inequality [81], which states that $\log \mathbb{E}[x] \geq \mathbb{E}[\log x]$, we can reformulate Equation (2.4) as follows:

$$\begin{aligned}\log P(A | \Theta) &= \log \sum_Z P(A, Z | \Theta) \\ &= \log \sum_Z q(Z) \frac{P(A, Z | \Theta)}{q(Z)} \\ &\geq \sum_Z q(Z) \log \frac{P(A, Z | \Theta)}{q(Z)} =: \mathcal{L}(q, \Theta).\end{aligned}\tag{2.5}$$

In this case, we define $x = \frac{P(A, Z | \Theta)}{q(Z)}$, and the expected value is computed with respect to the distribution q over the variables Z . Therefore, Equation (2.5) serves as a lower bound for the log-likelihood, and maximizing $\mathcal{L}(\Theta)$ is equivalent to maximizing $\mathcal{L}(q, \Theta)$.

During the E step of the algorithm, we use the current parameter values Θ^{old} to maximize the lower bound $\mathcal{L}(q, \Theta^{old})$ with respect to the distribution $q(Z)$. The solution to this maximization problem is

achieved when the lower bound matches the log-likelihood, and this equality is established when

$$\begin{aligned}
 q(Z) &= \frac{P(A, Z | \Theta^{old})}{\sum_Z P(A, Z | \Theta^{old})} \\
 &= \frac{P(A, Z | \Theta^{old})}{P(A | \Theta^{old})} \\
 &= P(Z | A, \Theta^{old}).
 \end{aligned} \tag{2.6}$$

Therefore, we can set $q(Z)$ as the posterior distribution of the unknown variables Z . In the M step, we keep $q(Z)$ fixed and maximize $\mathcal{L}(q, \Theta)$ with respect to Θ to obtain a new value Θ^{new} . By substituting $q(Z) = P(Z | A, \Theta^{old})$ into Equation (2.5), we observe that after the E step, the lower bound takes on the following form:

$$\begin{aligned}
 \mathcal{L}(q, \Theta) &= \sum_Z P(Z | A, \Theta^{old}) \log P(A, Z | \Theta) - \sum_Z P(Z | A, \Theta^{old}) \log P(Z | A, \Theta^{old}) \\
 &= Q(\Theta, \Theta^{old}) + \text{const},
 \end{aligned} \tag{2.7}$$

where the constant term corresponds to the negative entropy of the q distribution and is therefore independent of Θ . Hence, the quantity being maximized in the M step is the expectation of the log-likelihood $P(A, Z | \Theta)$ with respect to the posterior distribution of the unknown variables $P(Z | A, \Theta^{old})$ [19]. The EM algorithm can also be used to find MAP estimates for the parameters, and in such case, the M step seeks to maximize $Q(\Theta, \Theta^{old}) + \log P(\Theta)$. Following the M step, we set $\Theta^{old} = \Theta^{new}$ and continue iterating between the E and the M steps until the algorithm converges.

The EM algorithm typically benefits from closed-form updates in both the E and M steps, which enhances the efficiency and scalability of this approach. However, it is important to note that this algorithm converges to a local optimum and does not provide a guarantee of reaching the global optimum. In practice, it is common to run the algorithm multiple times using different initializations, which may result in the convergence of different local optima. Afterward, one can choose the best realization, that is, for instance, the one with the highest likelihood upon convergence.

2.2.3 Variational Inference

Variational Inference (VI) is a statistical technique used in probabilistic modeling to approximate posterior probability distributions [82, 150]. This method becomes particularly useful when calculating the exact posterior distribution of the latent variables in a probabilistic model is difficult or computationally unfeasible. The main idea behind VI is to choose an approximation from a tractable set of distributions, and then strives to minimize the discrepancy between this approximation and the actual posterior distribution. Thus, VI transforms the inference process into an optimization problem [21, 89, 104].

Following the Bayesian framework, we can define the posterior distribution of the latent variables Θ given the observed data A as follows:

$$P(\Theta | A) = \frac{P(A | \Theta)P(\Theta)}{P(A)} = \frac{P(A, \Theta)}{P(A)}. \tag{2.8}$$

In this setting, $P(A, \Theta)$ represents the joint distribution that characterizes the generative process underlying the observed data, while $P(A)$ corresponds to the model evidence. The computation

of this quantity involves the marginalization of the latent variables from the joint distribution, which requires the evaluation of $\int P(A, \Theta) d\Theta$ for continuous variables or $\sum_{\Theta} P(A, \Theta)$ for discrete variables. Nonetheless, performing any of these operations is often analytically intractable, making the derivation of $P(\Theta | A)$ unfeasible.

The main idea behind VI is to provide an approximation of the true and intractable posterior distribution $P(\Theta | A)$, with the objective of making this approximation as close as possible to the true posterior. To achieve this, we first define a family of approximate and manageable distributions, denoted as \mathcal{Q} , over the latent variables. Within this family, each $q(\Theta) \in \mathcal{Q}$ serves as a potential approximation to the exact posterior. VI then seeks to find the member of this family that is closest to the intractable posterior, where closeness is measured using the Kullback-Leibler (KL) divergence [92]:

$$\begin{aligned} \text{KL}(q(\Theta) || P(\Theta | A)) &= \int q(\Theta) \log \left\{ \frac{q(\Theta)}{P(\Theta | A)} \right\} d\Theta \\ &= \mathbb{E}_q[\log q(\Theta)] - \mathbb{E}_q[\log P(\Theta | A)] \\ &= \mathbb{E}_q[\log q(\Theta)] - \mathbb{E}_q[\log P(\Theta, A)] + \log P(A). \end{aligned} \quad (2.9)$$

Therefore, the optimal variational distribution $q^*(\Theta)$ among the set of distributions within the family \mathcal{Q} is obtained by solving the following optimization problem:

$$q^*(\Theta) = \arg \min_{q(\Theta) \in \mathcal{Q}} \text{KL}(q(\Theta) || P(\Theta | A)). \quad (2.10)$$

Nevertheless, this minimization problem remains unfeasible due to the presence of the model evidence $\log P(A)$ in the KL divergence, as demonstrated in Equation (2.9). To overcome this, VI optimizes an alternative objective function that is equivalent to the KL up to an added constant [21]. This quantity is known as the evidence lower bound (ELBO) and is defined as follows:

$$\text{ELBO} := \mathbb{E}_q[\log P(\Theta, A)] - \mathbb{E}_q[\log q(\Theta)]. \quad (2.11)$$

By definition, the ELBO corresponds to the negative KL divergence plus $\log P(A)$, which is a constant with respect to $q(\Theta)$. Therefore, minimizing the KL divergence between the variational distribution and the true posterior is equivalent to maximizing the ELBO [113].

To fully define the optimization problem, we need to specify a variational family \mathcal{Q} . One of the most commonly employed methods is based on the mean-field approximation [112]. This approach assumes that the latent variables are independent of each other, so that the joint distribution $q(\Theta)$ can be factorized into a product of individual distributions for each variable. Therefore, a mean-field variational distribution is represented as:

$$q(\Theta) = \prod_i q_i(\Theta_i), \quad (2.12)$$

where $q_i(\Theta_i)$ represents the individual approximating distributions for each latent variable Θ_i . Notably, this approach does not impose any specific parametric constraints on these distributions.

The factorization in Equation (2.12) significantly reduces the computational complexity of the inference process. Indeed, under the mean-field approximation, it is possible to get closed-form expressions for the updates of $q_i^*(\Theta_i)$ by applying the Coordinate Ascent Variational Inference (CAVI) algorithm [19]. CAVI is an iterative approach that optimizes each factor of the mean-field

variational distribution, while keeping the others fixed, until it reaches a local maximum of the ELBO. Following this approach, the optimal variational distributions are defined as:

$$q_i^*(\Theta_i) \propto \exp \left\{ \mathbb{E}_{-i} \left[\log P(\Theta_i | \Theta_{-i}, A) \right] \right\}. \quad (2.13)$$

In this equation, $P(\Theta_i | \Theta_{-i}, A)$ denotes the complete conditional of Θ_i , which is the conditional distribution of Θ_i given all the other latent variables and the observed data. The expectation $\mathbb{E}_{-i}[\cdot]$ is then taken with respect to the joint variational distribution over Θ_{-i} , that is, $\prod_{j \neq i} q_j(\Theta_j)$. An alternative result equivalent to Equation (2.13) states that, fixed the other variational distributions $q_j(\Theta_j)$, $j \neq i$, the optimal $q_i^*(\Theta_i)$ is proportional to:

$$q_i^*(\Theta_i) \propto \exp \left\{ \mathbb{E}_{-i} \left[\log P(\Theta_i, \Theta_{-i}, A) \right] \right\}. \quad (2.14)$$

This approach is quite versatile and can be employed to infer both discrete and continuous latent variables, using a variety of parametric forms for q_i . In particular, an important result is achieved when the complete conditional belongs to an exponential family. In such cases, each of the variational distributions also falls within the same exponential family as its corresponding complete conditional, and this result simplifies the derivation of the corresponding CAVI algorithm [21, 75]. To elaborate further, consider v_i as the variational parameter for the parametric distribution $q_i(\Theta_i)$. Then, in each update step of the CAVI algorithm, instead of setting $q_i^*(\Theta_i)$ as in Equation (2.13) or Equation (2.14), we simply set its parameter equal to the expected parameter of its complete conditional:

$$v_i = \mathbb{E} \left[\eta_i(\Theta_{-i}, A) \right]. \quad (2.15)$$

This formulation simplifies the development of CAVI algorithms for a wide range of complex models, making it feasible to perform approximate inference for many real-world datasets.

2.3 Community detection models

Real-world data exhibit remarkable complexity, often arising from interactions that are not purely random. In fact, many real-world networks display an inherent structural organization, which is characterized by a built-in community structure [40]. For instance, in social networks, individuals naturally cluster with their friends [68], while in citation networks, these communities might represent groups of related papers focused on specific topics [131]. The process of uncovering these hidden communities is commonly referred to as *community detection*, and it plays a crucial role in enhancing our comprehensive understanding of real-world networks.

Within this section, we delve into the notion of community structure, along with its main characteristics and representations. Furthermore, we expound two state-of-the-art latent variable models designed to infer communities from network data. Specifically, we introduce the pioneering Stochastic Block Model, which is a foundational framework for community detection, and the Multitensor model, an extension that accommodates overlapping communities and multilayer networks. This latter model also serves as the primary building block for the methods developed in this thesis.

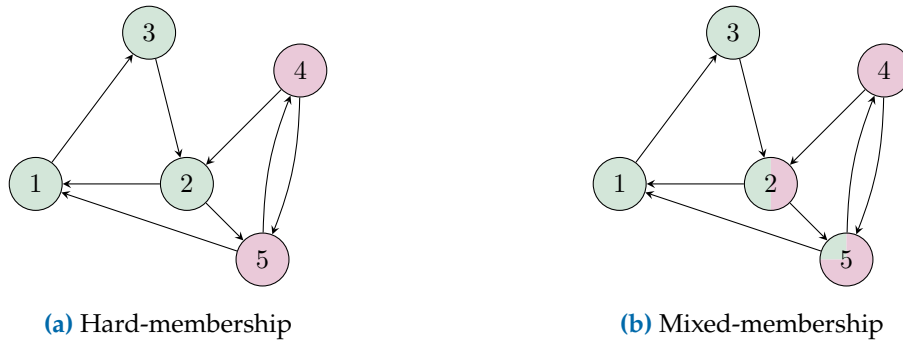


Figure 2.5: Illustration of two different community partitions. (a) The graphical representation of a hard-membership partition, where nodes are exclusively assigned to a single community, either pink or green. (b) The graphical representation of a mixed-membership partition, where nodes can belong to more than one group with different probabilities.

2.3.1 Community structures

Communities, also called *clusters* or *modules*, are groups of vertices within a graph that share common properties or perform similar roles [60]. The notion of communities is crucial in network analysis as it unveils hidden structures and functional units within networks. Additionally, it provides a more compact and lower-dimensional representation of complex systems, which is essential for analyzing large-scale networks. In particular, we focus on *overlapping communities*, a scenario where nodes can simultaneously belong to more than one group [106]. This approach provides a more accurate representation of real-world data, where nodes are expected to be part of multiple groups. For instance, in social media networks, users can engage in numerous communities based on their diverse interests. Similarly, in protein-protein interaction networks, proteins may participate in various biological pathways or functions, leading to their inclusion in multiple overlapping communities within the network.

Figure 2.5 visually illustrates these different scenarios. In Figure 2.5a, nodes are exclusively assigned to a single community, either pink or green, denoting a *hard-membership* partition. Conversely, in Figure 2.5b, the two communities overlap, and nodes have the flexibility to belong to both, representing a *mixed-membership* partition. This is the scenario underlying the models outlined in this thesis. For each node i , we measure the strength of its membership to the communities using a probability vector u_i of length K , where K represents the number of communities. In the case of hard-membership, these vectors contain only one non-zero entry. In contrast, in mixed-membership scenarios, nodes may belong to different communities with distinct probabilities. As an example, in Figure 2.5b, nodes 2 and 5 have their memberships represented by $u_2 = [0.5, 0.5]$ and $u_5 = [0.25, 0.75]$, respectively. Instead, the memberships of all other nodes are indicated as either $u_i = [1, 0]$ or $u_i = [0, 1]$.

Whether we consider hard- or mixed-membership partitions, the communities bring together nodes that exhibit similar connection patterns compared to nodes in other groups [119]. The overall community structure is then determined by the nature of the connections between the different communities. To capture and represent this information, community detection models introduce an *affinity matrix* w of dimension $K \times K$, where each entry w_{kq} represents the density of edges between each pair of groups. Real-world networks exhibit various types of community structures, and the affinity matrices of some of them are illustrated in Figure 2.6. Among these structures, one of the most prevalent is the *assortative structure*, often referred to as *homophily* in the context of social networks [102]. In this setting, communities represent groups of nodes that are

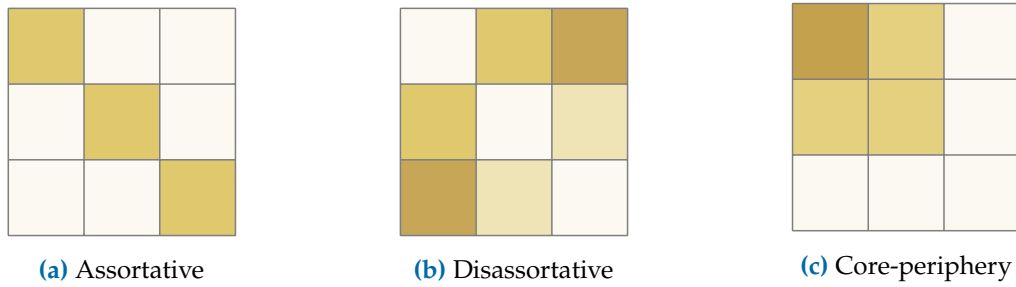


Figure 2.6: Illustration of three community structures. The graphical representation of the affinity matrices w for three different community structures. (a) In an assortative structure, nodes tend to connect more within their own communities rather than with nodes from other communities. (b) In a disassortative structure, edges are more likely to exist between groups than within them. (c) In a core-periphery structure, nodes in the core strongly connect among themselves, while the peripheral nodes are weakly connected.

densely interconnected, resulting in significantly higher edge densities within the diagonal blocks than between them, as depicted in Figure 2.6a. Conversely, Figure 2.6b illustrates a *disassortative structure*, where edges are more likely to exist between groups than within them. Additionally, some real-world data follow a different organizational pattern known as *core-periphery structure* [25], shown in Figure 2.6c. In such networks, nodes that strongly connect among themselves represent the core, while a separate periphery comprises weakly connected vertices.

In real-world data, these community structures are typically unknown and the aim of community detection algorithms is to unveil these hidden patterns, thereby offering a meaningful interpretation of real-world interactions. In the literature, several approaches exist for the detection of communities [61]. Classical methods encompass modularity optimization [108], spectral clustering [149], and hierarchical clustering [151]. Modularity optimization seeks to maximize the density of connections within communities while minimizing links between them. Spectral clustering, instead, employs the eigenvalues of the network’s laplacian matrix to partition nodes into distinct communities. On the other hand, hierarchical clustering builds a multi-level hierarchy of communities. Although these techniques are widely adopted and have significantly contributed to network analysis, they have limitations. For instance, they often rely exclusively on hard-membership partitions and assortative community structures. A broader and more flexible framework is that of probabilistic generative models, such as those we develop in this thesis. This approach is more powerful because it does not impose any structural constraint, enables various inference tasks, and accommodates the detection of overlapping communities. In the next subsections, we will introduce two state-of-the-art latent variable models that form the foundations for our methods.

2.3.2 Stochastic Block Model

The *Stochastic Block Model (SBM)* is arguably the simplest probabilistic generative model for graphs that exhibit community structures [1]. It was originally introduced by sociologists in 1983 [77], and it still serves as the foundational framework and benchmark for community detection models. In essence, the SBM posits that each node i belongs exclusively to a single community or block b_i , and the interactions among these nodes are entirely determined by their community memberships $b = (b_1, \dots, b_N)$. This fundamental assumption is based on the concept of *stochastic equivalence*, according to which if two nodes i and j belong to the same community k , they share the same probability of connecting with any other node h . The likelihood of interactions among these communities is regulated by the affinity matrix w , where each entry w_{kq} characterizes the

probability of an edge existing between a node in community k and one in community q . Thus, under the SBM assumptions, the probability of observing an interaction (i, j) can be expressed as:

$$P(A_{ij} | b, w) = w_{b_i b_j}, \quad (2.16)$$

and the likelihood for binary and undirected networks without self-loops is given by:

$$\begin{aligned} P(A | b, w) &= \prod_{i < j} P(A_{ij} | b, w) \\ &= \prod_{i < j} \text{Ber}(A_{ij}; p_{ij}) \\ &= \prod_{i < j} w_{b_i b_j}^{A_{ij}} [1 - w_{b_i b_j}]^{1 - A_{ij}}. \end{aligned} \quad (2.17)$$

It is worth noting that when all entries of the affinity matrix w are uniform and set to a constant value p , the SBM is equivalent to the Erdős–Rényi model. Conversely, when the matrix w features diverse entries, for instance encoding one of the structures depicted in [Figure 2.6](#), the SBM has the capability to generate networks with a planted partition b .

The generative model outlined in [Equation \(2.17\)](#) represents the most basic form of the SBM, and numerous extensions have been developed over the years. For example, there are variations tailored to handle weighted networks [84], to accommodate mixed-membership partitions [2], and to address dynamic networks [154], among others. Regardless of the specific method considered, a common challenge is to infer the latent variables of these models from the available data. In the literature, a diverse range of methodologies has been employed to address this inference task. For instance, to get MLEs for the parameters, one can opt for either greedy algorithms with local optimal moves, or employ EM techniques [45, 84, 143]. Alternatively, within a Bayesian formalism, one can apply Markov chain Monte Carlo methods [109, 111, 122] or VI methods [2, 70, 86]. Additionally, another viable approach involves utilizing variants of the belief propagation or message passing algorithms [46, 47, 64]. This list is not exhaustive, and for a more comprehensive discussion, interested readers can refer to review articles [62, 95].

Another relevant issue for this class of models is the selection of the number K of communities, which inherently represents a model selection problem. Also in this context, several strategies have emerged in the literature. Traditional approaches to address this challenge encompass the computation of statistical metrics such as the Akaike information criterion [3] or the Bayesian information criterion [140]. Nevertheless, in the realm of network modeling, a more prevalent methodology involves considering the minimum description length principle [120, 121, 123]. Additionally, other criteria include the use of the integrated complete likelihood [36, 43, 100] and the application of variational Bayesian approaches [76, 109, 124]. However, in the models presented in this thesis, we use another procedure and we select K following a cross-validation routine [32, 45]. In this approach, the dataset is divided into training and test sets: the training set is employed to fit the model and infer the parameters, while the test set is used to evaluate the model's performance given the inferred parameters. For instance, one can predict the edges within the hidden test set and compute the Area Under the Curve (AUC) [73] to assess the model's ability to retrieve such information. This procedure is then iterated for different training and test partitions, and it is also repeated for various values of K . Afterward, the optimal K is chosen based on the best performance observed across all iterations in the test sets. To find more information, interested readers can refer to the publications in [Appendix A](#).

2.3.3 Multitensor

Multitensor [45] is a probabilistic generative model designed to perform inference and community detection in multilayer networks. This model is versatile, as it can handle both directed and undirected multilayer networks, as well as nonnegative discrete weights. The core assumption of Multitensor is the existence of overlapping communities shared across all network layers. Although the partition is common to all layers, this model provides the flexibility for each layer to exhibit distinct connectivity patterns, including arbitrarily mixtures of assortative, disassortative and core-periphery structures. Formally, Multitensor assigns two K -dimensional mixed-membership vectors to each node i , denoted as u_i and v_i , which respectively represent its outgoing and incoming communities memberships. In the case of undirected networks, the model sets these vectors to be equal, i.e., $u = v$. In addition, Multitensor characterizes the community structure of the entire multilayer network with an affinity tensor $w = \{w_{kq}^\ell\} \in \mathbb{R}^{L \times K \times K}$, wherein each affinity matrix w^ℓ describes the structure of a specific layer ℓ , which may differ from one layer to another. Under these assumptions, the expected number of edges from node i to node j in layer ℓ is expressed with the following bilinear form:

$$\lambda_{ij}^\ell = \sum_{k,q=1}^K u_{ik} w_{kq}^\ell v_{jq}. \quad (2.18)$$

Multitensor assumes that the edges of a weighted multilayer network $A = \{A_{ij}^\ell\} \in \mathbb{N}_0^{L \times N \times N}$ are conditionally independent given the parameters, and it models the likelihood of the data as follows:

$$\begin{aligned} P(A | u, v, w) &= \prod_{\ell=1}^L \prod_{i,j=1}^N P(A_{ij}^\ell | u, v, w) \\ &= \prod_{\ell=1}^L \prod_{i,j=1}^N \text{Pois}(A_{ij}^\ell; \lambda_{ij}^\ell) \\ &= \prod_{\ell=1}^L \prod_{i,j=1}^N \frac{e^{-\lambda_{ij}^\ell} (\lambda_{ij}^\ell)^{A_{ij}^\ell}}{A_{ij}^\ell!}. \end{aligned} \quad (2.19)$$

Given an observed network, the goal is to simultaneously infer the nodes' membership vectors u and v , and the affinity matrices w^ℓ for each layer. To perform this task, the authors employed an efficient and highly-scalable EM algorithm. Furthermore, they evaluated the model's performance on a variety of synthetic and real-world networks, focusing not only on tasks such as community detection but also on link prediction. Additionally, they introduced a principled approach to quantify the interdependence between the layers within a multilayer network, thereby facilitating the identification of redundant or highly independent layers.

The flexibility of this method, together with its ability to capture the complexities of real-world data, makes it a versatile and robust foundation. In the models presented in this thesis, we extend and adapt the underlying assumptions of Multitensor to accommodate attributed multilayer networks, dynamic networks, and hypergraphs. Additionally, we incorporate other mechanisms like reciprocity, which represents the tendency of a pair of nodes to form mutual connections between each other.

3 Published Work

This thesis is based on ten peer-reviewed publications, where I am either first or second author. We group these work in three sections, that reflect the three principal directions investigated during the doctoral studies. This chapter is thus divided as follows:

- ▶ In [Section 3.1](#), we present publications that delve into the analysis of attributed multilayer networks. This section encompasses two peer-reviewed papers. Among these work, one introduces a new generative model and its properties to perform inference on such data, while the second one is focused on demonstrating the effectiveness of these methodologies in the context of patent citation networks.
- ▶ In [Section 3.2](#), we discuss methods to handle the pairwise dependencies between two directed edges connecting node pairs, that comes with the relaxation of the conditional independence assumption in network models. Within this section, we have compiled a total of five peer-reviewed papers. Two of these work offer a theoretical perspective on the subject, while the remaining three extend this framework to other contexts, such as dynamic networks, anomaly detection, and multiply-reported data.
- ▶ In [Section 3.3](#), we describe techniques to characterize the structural organization of higher-order data. This section includes three peer-reviewed publications. Two of these present mathematical approaches for performing inference on hypergraphs, while the third work introduces a Python library that offers a broad range of tools and algorithms to handle data with higher-order interactions.

Every section begins with an introduction that explains the relationship between the various publications and their relevance to the primary research direction. Following this preamble, each paper is accompanied by an abstract, an explanation of author contributions, and a summary of the publication venue. All the publications referenced in the sections can be found in [Appendix A](#) for easy access.

3.1 Inference on attributed multilayer networks

In this section, we introduce various techniques that push on the analysis of attributed multilayer networks. These are complex network representations that describe multiple types of interactions among the same set of units (nodes), while also incorporating node information such as attributes or covariates. Consequently, the approaches discussed here combine conveniently and in a principled way various sources of information, leveraging both the network topology and the node metadata.

The first publication introduces MTCOV, a probabilistic generative model designed to perform community detection and broader inference in attributed multilayer networks. This approach relies on a linear combination of the multilayer structure and the node information, and offers a tool to quantitatively measure the influence of the node attributes given in input.

The second publication demonstrates the application of MTCOV in the analysis of patent citation networks, with the aim to comprehensively comprehend the worldwide technological landscape

that patent data can provide. For the first time, we analyze these networks employing a multilayer framework that can incorporate contextual and jurisdictional details from various patent citations.

3.1.1 Community detection with node attributes in multilayer networks

Abstract Community detection in networks is commonly performed using information about interactions between nodes. Recent advances have been made to incorporate multiple types of interactions, thus generalizing standard methods to multilayer networks. Often, though, one can access additional information regarding individual nodes, attributes, or covariates. A relevant question is thus how to properly incorporate this extra information in such frameworks. Here we develop a method that incorporates both the topology of interactions and node attributes to extract communities in multilayer networks. We propose a principled probabilistic method that does not assume any a priori correlation structure between attributes and communities but rather infers this from data. This leads to an efficient algorithmic implementation that exploits the sparsity of the dataset and can be used to perform several inference tasks; we provide an open-source implementation of the code online. We demonstrate our method on both synthetic and real-world data and compare performance with methods that do not use any attribute information. We find that including node information helps in predicting missing links or attributes. It also leads to more interpretable community structures and allows the quantification of the impact of the node attributes given in input.

Author contribution In this project, I took on the role of the first author. Together with my advisor, we conceived the research and designed the experiments. Additionally, I was responsible for implementing the model, writing and refining the code. I also analyzed the data together with the team to uncover significant insights from the results, and I played a crucial part in creating a variety of different visualizations. Collaboratively with my co-authors, I actively participated in co-writing the manuscript. Additionally, I made significant contributions during the rebuttal phase by conducting supplementary experiments, addressing reviewers' feedback, and editing the manuscript.

Venue *Scientific Reports* is a peer-reviewed and open access journal published by Nature Portfolio since 2011. It covers original research from across all areas of the natural sciences, psychology, medicine, and engineering. The primary aim of this journal is to assess the scientific rigor of submitted papers, emphasizing their validity rather than subjective importance or impact. By adopting an open access policy, this journal offers researchers high visibility for their work. Notably, in September 2016, *Scientific Reports* became the largest journal in the world in terms of the number of published articles, and it holds the position of the 5th most-cited journal worldwide.

3.1.2 Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships

Abstract The use of patent citation networks as research tools is becoming increasingly commonplace in the field of innovation studies. However, these networks rarely consider the contexts in which these citations are generated and are generally restricted to a single jurisdiction. Here, we propose and explore the use of a multilayer network framework that can naturally incorporate

citation metadata and stretch across jurisdictions, allowing for a complete view of the global technological landscape that is accessible through patent data. Taking a conservative approach that links citation network layers through triadic patent families, we first observe that these layers contain complementary, rather than redundant, information about technological relationships. To probe the nature of this complementarity, we extract network communities from both the multilayer network and analogous single-layer networks, then directly compare their technological composition with established technological similarity networks. We find that while technologies are more splintered across communities in the multilayer case, the extracted communities match much more closely the established networks. We conclude that by capturing citation context, a multilayer representation of patent citation networks is, conceptually and empirically, better able to capture the significant nuance that exists in real technological relationships when compared to traditional, single-layer approaches. We suggest future avenues of research that take advantage of novel computational tools designed for use with multilayer networks.

Author contribution As the second author of this work, my contributions involved providing support in conceptualizing the experiments and applying the methodology. Specifically, I played an important role in pre-processing the dataset, optimizing the code to handle the large dataset efficiently, and conducting the experiments. Alongside the co-authors, we collectively validated the results, collaborated to create clear and informative visualizations, and participated in the writing of the manuscript. Additionally, during the rebuttal phase, I contributed to refining both the response to the reviewers and the revised manuscript.

Venue *Technological Forecasting & Social Change* (TFSC) is a peer-reviewed journal published by Elsevier. Since 1969, it focuses on research at the intersection of technology, innovation, and societal change. The TFSC journal publishes articles that explore various aspects of technological forecasting, including the analysis of emerging technologies, innovation diffusion, adoption patterns, and the implications of technological changes for different sectors of society. It also examines the social, economic, political, and environmental consequences of technological developments.

3.2 Reciprocity and the relaxation of the conditional independence assumption

This section presents methods that contribute to the development of approaches to incorporate reciprocity in the analysis of directed networks. Reciprocity is the tendency of a pair of nodes to form mutual connections between each other. Therefore, the methods presented here differ from standard generative models in that they can handle the interdependence between two directed edges that connect pairs of nodes. Mathematically, this means that we jointly model the edges involving the same node pairs (i, j) , denoted as $P(A_{ij}, A_{ji} | \Theta)$, instead of the usual practice of considering $P(A_{ij} | \Theta)$ and $P(A_{ji} | \Theta)$ independently from each other.

The first two publications investigate theoretical perspectives on the subject, each taking a different approach that serves as a foundational component of this research direction. The two methods, CRep and JointCRep, differ in their generative process. CRep models conditional distributions with Poisson distributions and relies on a pseudo-likelihood approximation, while JointCRep utilizes the

properties of the Bivariate-Bernoulli distribution to model the joint distributions of edges involving the same pairs of nodes in closed-form.

The third and the fourth work build upon the aforementioned mathematical frameworks and apply them to different contexts. Specifically, the third publication extends CRep to analyze dynamic networks, which are networks that change over time. In the fourth paper, instead, the formalism of JointCRep is employed to develop a probabilistic generative approach that can be used to perform anomaly detection on the edges of a network.

The fifth publication presents a method to estimate the unobserved network structure from multiply reported data. In this model, the reciprocity parameter is incorporated by following the principles of CRep, and reflects the intuition that reporters tend to nominate the same individuals for both directions of a relationship.

3.2.1 Generative model for reciprocity and community detection in networks

Abstract We present a probabilistic generative model and efficient algorithm to model reciprocity in directed networks. Unlike other methods that address this problem such as exponential random graphs, it assigns latent variables as community memberships to nodes and a reciprocity parameter to the whole network rather than fitting order statistics. It formalizes the assumption that a directed interaction is more likely to occur if an individual has already observed an interaction towards her. It provides a natural framework for relaxing the common assumption in network generative models of conditional independence between edges, and it can be used to perform inference tasks such as predicting the existence of an edge given the observation of an edge in the reverse direction. Inference is performed using an efficient expectation-maximization algorithm that exploits the sparsity of the network, leading to an efficient and scalable implementation. We illustrate these findings by analyzing synthetic and real data, including social networks, academic citations and the Erasmus student exchange program. Our method outperforms others in both predicting edges and generating networks that reflect the reciprocity values observed in real data, while at the same time inferring an underlying community structure. We provide an open-source implementation of the code online.

Author contribution For this project, I shared the co-authorship with a colleague of mine. One of my primary responsibilities was to analyze and verify the mathematical details to support the model's design and implementation. Moreover, I contributed to the code implementation, experiment design and results visualization. During the experimentation phase, my specific task was to conduct experiments using synthetic data. Furthermore, I worked closely with my colleague in drafting and editing the manuscript to ensure that the research findings were presented precisely and clearly. Lastly, I actively participated in the two rounds of the rebuttal process, working together with my colleague to address all the points raised by the reviewers and revise the manuscript accordingly.

Venue *Physical Review Research* (PRR) is a fully open access, peer-reviewed, and multidisciplinary journal that was launched in 2019. It is part of the Physical Review family and published by the American Physical Society (APS). PRR welcomes papers covering a wide range of research topics that are relevant to the field of physics. The journal's scope includes both fundamental and

applied research, as well as theoretical and experimental studies that incorporate technical and methodological innovations. Additionally, PRR is interested in interdisciplinary and emerging areas of research.

3.2.2 Community detection and reciprocity in networks by jointly modelling pairs of edges

Abstract To unravel the driving patterns of networks, the most popular models rely on community detection algorithms. However, these approaches are generally unable to reproduce the structural features of the network. Therefore, attempts are always made to develop models that incorporate these network properties beside the community structure. In this work, we present a probabilistic generative model and an efficient algorithm to both perform community detection and capture reciprocity in networks. Our approach jointly models pairs of edges with exact 2-edge joint distributions. In addition, it provides closed-form analytical expressions for both marginal and conditional distributions. We validate our model on synthetic data in recovering communities, edge prediction tasks, and generating synthetic networks that replicate the reciprocity values observed in real networks. We also highlight these findings on two real datasets that are relevant for social scientists and behavioral ecologists. Our method overcomes the limitations of both standard algorithms and recent models that incorporate reciprocity through a pseudo-likelihood approximation. The inference of the model parameters is implemented by the efficient and scalable expectation-maximization algorithm, as it exploits the sparsity of the dataset. We provide an open-source implementation of the code online.

Author contribution As the first author of this project, I was responsible for designing the model and planning the experiments, ensuring that they were aligned with the research objectives. Additionally, I was actively involved in the implementation of the model, writing and refining the code to ensure its accuracy and efficiency. I also led the analysis of the data, working collaboratively with the team to extract significant insights from the results and create meaningful visualizations. Furthermore, I wrote the majority of the manuscript, and coordinated the two rounds of the project's rebuttal.

Venue *Journal of Complex Networks* is a peer-reviewed journal published by Oxford University Press that was established in 2013. The journal publishes articles and reviews that make a significant contribution to the analysis and understanding of complex networks and their applications in various fields. Its coverage ranges from the fundamental mathematical, physical, and computational principles required for studying complex networks to their practical applications, resulting in predictive models in diverse systems such as molecular, biological, ecological, informational, engineering, social, technological, and others.

3.2.3 Reciprocity, community detection, and link prediction in dynamic networks

Abstract Many complex systems change their structure over time, in these cases dynamic networks can provide a richer representation of such phenomena. As a consequence, many inference methods have been generalized to the dynamic case with the aim to model dynamic interactions. Particular interest has been devoted to extend the stochastic block model and its variant, to capture community

structure as the network changes in time. While these models assume that edge formation depends only on the community memberships, recent work for static networks show the importance to include additional parameters capturing structural properties, as reciprocity for instance. Remarkably, these models are capable of generating more realistic network representations than those that only consider community membership. To this aim, we present a probabilistic generative model with hidden variables that integrates reciprocity and communities as structural information of networks that evolve in time. The model assumes a fundamental order in observing reciprocal data, that is an edge is observed, conditional on its reciprocated edge in the past. We deploy a Markovian approach to construct the network's transition matrix between time steps and parameters' inference is performed with an expectation-maximization algorithm that leads to high computational efficiency because it exploits the sparsity of the dataset. We test the performance of the model on synthetic dynamical networks, as well as on real networks of citations and email datasets. We show that our model captures the reciprocity of real networks better than standard models with only community structure, while performing well at link prediction tasks.

Author contribution In my role as the second author of this work, I provided support in analyzing the data, working closely with the first author to accurately interpret the findings. I was also responsible to proofread the manuscript and ensure that it was free of errors and coherent in its presentation. Additionally, I contributed to the three rounds of the rebuttal process, assisting in addressing any concerns or questions raised by the reviewers.

Venue *Journal of Physics: Complexity* is a new fully open access, peer-reviewed, and interdisciplinary journal published by IOP Publishing. Launched in 2020, the journal's objective is to publish high-quality quantitative research in the field of complexity. It aims to present important scientific advancements in theoretical, experimental, and applied physics-related research that enhance our scientific knowledge of complex systems and networks.

3.2.4 Anomaly, reciprocity, and community detection in networks

Abstract Anomaly detection algorithms are a valuable tool in network science for identifying unusual patterns in a network. These algorithms have numerous practical applications, including detecting fraud, identifying network security threats, and uncovering significant interactions within a data set. In this project, we propose a probabilistic generative approach that incorporates community membership and reciprocity as key factors driving regular behavior in a network, which can be used to identify potential anomalies that deviate from expected patterns. We model pairs of edges in a network with exact two-edge joint distributions. As a result, our approach captures the exact relationship between pairs of edges and provides a more comprehensive view of social networks. Additionally, our study highlights the role of reciprocity in network analysis and can inform the design of future models and algorithms. We also develop an efficient algorithmic implementation that takes advantage of the sparsity of the network.

Author contribution In this project, I held the role of the second author and my main contribution involved providing support during the model design phase. Specifically, I played a key role in validating the model equations to ensure their accuracy and reliability. Additionally, I thoroughly proofread the manuscript and offered assistance during the editing process. Furthermore, I

contributed to the two rounds of the rebuttal phase by refining both the response to the reviewers and the revised manuscript.

Venue *Physical Review Research* (PRR) is a fully open access, peer-reviewed, and multidisciplinary journal that was launched in 2019. It is part of the Physical Review family and published by the American Physical Society (APS). PRR welcomes papers covering a wide range of research topics that are relevant to the field of physics. The journal's scope includes both fundamental and applied research, as well as theoretical and experimental studies that incorporate technical and methodological innovations. Additionally, PRR is interested in interdisciplinary and emerging areas of research.

3.2.5 Latent network models to account for noisy, multiply reported social network data

Abstract Social network data are often constructed by incorporating reports from multiple individuals. However, it is not obvious how to reconcile discordant responses from individuals. There may be particular risks with multiply reported data if people's responses reflect normative expectations—such as an expectation of balanced, reciprocal relationships. Here, we propose a probabilistic model that incorporates ties reported by multiple individuals to estimate the unobserved network structure. In addition to estimating a parameter for each reporter that is related to their tendency of over- or under-reporting relationships, the model explicitly incorporates a term for "mutuality", the tendency to report ties in both directions involving the same alter. Our model's algorithmic implementation is based on variational inference, which makes it efficient and scalable to large systems. We apply our model to data from a Nicaraguan community collected with a roster-based design and 75 Indian villages collected with a name-generator design. We observe strong evidence of "mutuality" in both datasets, and find that this value varies by relationship type. Consequently, our model estimates networks with reciprocity values that are substantially different than those resulting from standard deterministic aggregation approaches, demonstrating the need to consider such issues when gathering, constructing, and analysing survey-based network data.

Author contribution On this project, I was the second author among a large team of researchers. I contributed to the model design and assisted with the validation of the mathematical derivations. Moreover, I collaborated closely with a few team members to develop the code, plan the experiments, and coordinate the data analysis. I primarily focused on analyzing synthetic data and assisting with results visualizations. Furthermore, I was actively involved in the draft process of the manuscript. Specifically, I took responsibility for writing the section on synthetic experiments and portions of the appendix. Additionally, I proofread the entire manuscript to ensure coherence and flow. During the rebuttal phase, I helped by running additional experiments and further correcting the manuscript.

Venue *Journal of the Royal Statistical Society Series A: Statistics in Society* is a peer-reviewed journal of statistics published by Oxford University Press for the Royal Statistical Society. Since 1988, the journal has been publishing high-quality papers that showcase the importance of statistical thinking, design, and analysis in various fields, without any subject-matter restrictions. Its emphasis is on well-written and clearly reasoned quantitative approaches to real-world problems, rather than technical details. Of particular interest are papers on topical or controversial statistical issues,

reviews of current statistical concerns, and those that demonstrate how statistical thinking has contributed to our understanding of important questions.

3.3 Community detection and the analysis of higher-order data

Within this section, we present methodologies that expand the set of statistical inference techniques available for analyzing higher-order data. These represent systems which involve interactions between groups of nodes of any arbitrary size. As a result, the models discussed here extend beyond conventional dyadic interactions (represented as A_{ij}) and instead focus on hyperedges, referred to as d -dimensional interactions A_e , where $d = |e|$.

The first two publications introduce distinct frameworks aimed at characterizing the structural organization of hypergraphs. Both methods adopt a mixed-membership structure as generative process, and utilize Poisson distributions to model the hyperedges. Nevertheless, they diverge in their approach to defining the mean of the marginal distributions. Specifically, Hypergraph-MT describes a hyperedge by considering the product of the node memberships belonging to it, whereas Hy-MMSBM employs a bilinear form for capturing dependencies within the hyperedges.

The third work is computational and introduces the Python library `hypergraphx`, which is openly available and serves as a valuable resource for analyzing networked systems with higher-order interactions. This library offers a broad range of tools and algorithms for constructing, visualizing, and analyzing data with higher-order interactions.

3.3.1 Inference of hyperedges and overlapping communities in hypergraphs

Abstract Hypergraphs, encoding structured interactions among any number of system units, have recently proven a successful tool to describe many real-world biological and social networks. Here we propose a framework based on statistical inference to characterize the structural organization of hypergraphs. The method allows to infer missing hyperedges of any size in a principled way, and to jointly detect overlapping communities in presence of higher-order interactions. Furthermore, our model has an efficient numerical implementation, and it runs faster than dyadic algorithms on pairwise records projected from higher-order data. We apply our method to a variety of real-world systems, showing strong performance in hyperedge prediction tasks, detecting communities well aligned with the information carried by interactions, and robustness against addition of noisy hyperedges. Our approach illustrates the fundamental advantages of a hypergraph probabilistic model when modeling relational systems with higher-order interactions.

Author contribution I had the privilege of being the first author of this project. Working closely with my advisor, we conceptualized the model and planned the experiments, ensuring they effectively showcased the strengths of our algorithm. We also collaborated to get a computationally efficient implementation of our method. Moreover, I actively participated in the data analysis process and played a key role in generating insightful visualizations. Together with my co-authors, I co-wrote the manuscript, ensuring the effective delivery of our findings with a coherent structure and smooth flow. In addition, I took the lead in managing the two rounds of the rebuttal process. This involved conducting supplementary experiments, addressing reviewers' comments, and editing the manuscript.

Venue *Nature Communications* is a peer-reviewed and open access journal published by Nature Portfolio since 2010. This multidisciplinary journal covers a wide range of natural sciences, encompassing fields such as physics, chemistry, earth sciences, medicine, and biology. The journal is committed to publishing impactful papers that represent significant advancements in their respective domains. It has a global reach and is known for its high impact. Nature Communications aims to provide a platform for interdisciplinary research and encourages collaboration among scientists.

3.3.2 Community detection in large hypergraphs

Abstract Hypergraphs, describing networks where interactions take place among any number of units, are a natural tool to model many real-world social and biological systems. In this work we propose a principled framework to model the organization of higher-order data. Our approach recovers community structure with accuracy exceeding that of currently available state-of-the-art algorithms, as tested in synthetic benchmarks with both hard and overlapping ground-truth partitions. Our model is flexible and allows capturing both assortative and disassortative community structures. Moreover, our method scales orders of magnitude faster than competing algorithms, making it suitable for the analysis of very large hypergraphs, containing millions of nodes and interactions among thousands of nodes. Our work constitutes a practical and general tool for hypergraph analysis, broadening our understanding of the organization of real-world higher-order systems.

Author contribution In this project, my role was that of the second author. I provided valuable support in the planning of innovative and meaningful experiments. Additionally, I played a key role in creating impactful visualizations that effectively conveyed the results. Furthermore, I actively contributed to the editing and proofreading of the manuscript. Throughout the two rounds of the rebuttal phase, I assisted in rewriting specific sections of the manuscript, rerunning experiments, refining visualizations, as well as drafting and revising the responses to the reviewers.

Venue *Science Advances* is a fully open access, peer-reviewed, and multidisciplinary journal that was established in 2015. It is published by the American Association for the Advancement of Science (AAAS), the world's oldest and largest general science organization. Science Advances focuses on disseminating high-impact research papers and reviews spanning all areas of science, including both specific disciplines and broader interdisciplinary subjects. The journal's mission is to identify and promote significant advancements in science and engineering across a wide range of areas, while also offering readers a vetted selection of research.

3.3.3 Hypergraphx: a library for higher-order network analysis

Abstract From social to biological systems, many real-world systems are characterized by higher-order, non-dyadic interactions. Such systems are conveniently described by hypergraphs, where hyperedges encode interactions among an arbitrary number of units. Here, we present an open-source python library, hypergraphx (HGX), providing a comprehensive collection of algorithms and functions for the analysis of higher-order networks. These include different ways to convert data across distinct higher-order representations, a large variety of measures of higher-order organization

at the local and the mesoscale, statistical filters to sparsify higher-order data, a wide array of static and dynamic generative models, and an implementation of different dynamical processes with higher-order interactions. Our computational framework is general, and allows to analyse hypergraphs with weighted, directed, signed, temporal and multiplex group interactions. We provide visual insights on higher-order data through a variety of different visualization tools. We accompany our code with an extended higher-order data repository and demonstrate the ability of HGX to analyse real-world systems through a systematic analysis of a social network with higher-order interactions. The library is conceived as an evolving, community-based effort, which will further extend its functionalities over the years. Our software is available at <https://github.com/HGX-Team/hypergraphx>.

Author contribution In this project, I held the position of the second author within a large team of researchers. My contribution primarily focused on the implementation of two community detection algorithms and the development of a tutorial to demonstrate their usage. Additionally, I assisted in reformatting the overall package. Furthermore, I took the lead in analyzing the dataset we used as a case study, developing functions for visualizing the communities, and refining the final version of the main figure. Lastly, I contributed to the writing process for the "mesoscale structures" and "visualization" sections.

Venue *Journal of Complex Networks* is a peer-reviewed journal published by Oxford University Press that was established in 2013. The journal publishes articles and reviews that make a significant contribution to the analysis and understanding of complex networks and their applications in various fields. Its coverage ranges from the fundamental mathematical, physical, and computational principles required for studying complex networks to their practical applications, resulting in predictive models in diverse systems such as molecular, biological, ecological, informational, engineering, social, technological, and others.

4 Discussion & Conclusion

In this chapter, we provide a comprehensive overview of the three main research directions investigated within this thesis. For each of these topics, we highlight the key findings in relation to the outlined publications, and discuss potential future directions. Finally, we summarize the overarching contribution of this thesis, and delve into the broader implications of our research.

4.1 Inference on attributed multilayer networks

Advancements in data collection techniques have led to the acquisition of more comprehensive data, particularly by gathering additional information that characterizes the nodes and their interactions within real-world systems. These enriched data are effectively represented by attributed single-layer and multilayer networks, as described in [Section 2.1](#), and analyzing these networks can result in a more profound understanding of real-world data. Recent studies, primarily focusing on attributed single-layer networks [80, 107, 145, 153], have demonstrated that considering node attributes in network models can significantly enhance network inference, by for instance boosting prediction performance. Furthermore, exploring the interplay between edge structure and node metadata can yield valuable insights into the underlying organization and functional relationships within the network. Building upon these results as motivation, in this thesis, we took a further step by examining attributed multilayer networks, instead of single-layer networks, which arguably offer a more nuanced representation of real-world data. In doing so, we provided principled methods for combining node and edge metadata, along with strategies for their validation, moving beyond the common practice of aggregating the layers and applying single-layer techniques [33, 51, 116].

In [Section 3.1.1](#), we introduced MTCOV, a probabilistic generative model designed to effectively integrate node information and network topology to perform inference on attributed multilayer networks [38]. At its core, MTCOV posits the existence of a hidden mixed-membership partitioning of the nodes, and it assumes this as the underlying mechanism for determining both interactions and node covariates. Furthermore, it combines the likelihoods of these two sources of information – network topology and node metadata – through a linear combination, and employs an efficient EM algorithm to infer the latent variables representing the community memberships and their connectivity structures. Our model is flexible, as it can be applied to a variety of network datasets, whether directed, weighted, or multilayer, making it a valuable tool for analyzing data across diverse fields. As an example, in the paper introducing the model, we conducted an extensive study of social support networks [38]. Additionally, in [Section 3.1.2](#), we applied this methodology for the first time in the analysis of patent citation networks [74]. In this context, we not only illustrated the importance of using a multilayer framework for patent citation data analysis but also emphasized the role of a node covariate in driving the inference, alongside the structural information embedded within the network.

In our work on MTCOV, we illustrated that effectively integrating node attributes with topological information can lead to substantial enhancements in network inference, even in the context of attributed multilayer networks. These improvements have been showcased in tasks like community detection and prediction, where our model consistently outperformed approaches that exclusively rely on network structure. Moreover, we demonstrated that taking node information into account can

yield more robust results, especially when dealing with incomplete or imbalanced data. Specifically, we observed better outcomes when the node metadata are more informative and exhibit some degree of correlation with the information conveyed by the interactions. Subsequently, our findings have been reinforced by other researchers who investigated the impact of node attributes in the inference process. They found that attributes can enhance the inference only when all terms in the likelihoods or posteriors are of comparable magnitude or when attributes are perfectly correlated with the interactions [58].

Importantly, a property of MTCOV is that it does not make any prior assumptions regarding the importance of a node attribute. Instead, it directly assesses the attributes' impact in the inference process using a cross-validation technique. In our experiments, we showcased the flexibility of MTCOV in exploiting the attributes that exhibit higher informativeness while ignoring those with weaker correlations to the network structure. For example, we observed that the attribute "caste" is the most informative when analyzing social support networks [38], while the country of the priority office where patents are filed plays a significant role in the analysis of the patent citation networks, enabling a more accurate quantification of certain citation patterns [74]. When the node metadata offer valuable insights, MTCOV identifies communities that align with this information. This approach leads to more interpretable results, where attributes actively influence the inference process, rather than only serving as a posterior tool for evaluating node partitions – a practice that can potentially result in erroneous scientific conclusions [118].

MTCOV represents one of the first and few probabilistic models designed to perform inference on attributed multilayer networks. Nevertheless, this area of research still remains largely unexplored, offering numerous avenues for further development. As an example, an intriguing direction of future research could involve integrating a mixture of heterogeneous attributes, rather than just considering categorical covariates, as done in our model. In particular, it would be interesting to investigate systematic methods for combining these covariates in a principled manner, going beyond the approaches taken in other methodologies, like spectral embedding [152], deep learning [28], or differential evolution [127]. Typically, these methods combine multiple attributes in a deterministic way into a similarity matrix, which restricts our ability to comprehend the underlying data generation process and quantify associated uncertainties. Furthermore, MTCOV could be extended to accommodate layers with different data types. In fact, many existing models for analyzing multilayer networks assume that all layers share the same underlying generative process. However, this simplification does not fully capture the complexities of real-world systems, where diverse types of interactions coexist. For instance, in social networks, nodes can be connected in various ways, including binary relationships like friendships, nonnegative discrete interactions such as call counts, and continuous real-valued measurements like distances between their residences. Additionally, nodes may possess different attributes. These extensions would then create a general framework capable of not only representing the complexities of real-world data in a principled manner, but also assessing the relevance of node and edge metadata in the inference process, exploiting only the most informative information. However, delving into this research direction poses some challenges: such as the need to develop techniques to automatically rescale the data and to effectively leverage the diverse sources of information, especially when they vary in size. Additionally, it requires efforts to ensure efficient inference despite the increased volume of information. Lastly, an interesting prospect for future research would be to take these ideas forward to analyze data with additional types of information, such as time-varying networks, further expanding the toolbox of techniques available for modeling complex networks.

4.2 Reciprocity and the relaxation of the conditional independence assumption

Directed networks serve as a mathematical representation of real-world data in which interactions have a directionality, such as citations between scientific papers. To uncover the underlying patterns governing these networks, many existing models rely on community detection algorithms, which have the capability to reveal intriguing insights regarding the inherent structure of various real-world datasets [99]. As elaborated in Section 2.3, these models posit that the interactions are fully determined by some hidden partition of the nodes. While this assumption has demonstrated its reliability, alternative mechanisms can be considered when analyzing directed networks. For example, many real networks exhibit reciprocity, the tendency of a pair of nodes to form mutual connections between each other. This property can also provide new insights into the topology of real-world data [63], and we are keen on incorporating it into the statistical models used for network analysis. Some existing methods have already integrated reciprocity into their framework, such as exponential random graphs models [78, 98, 132] and stochastic oriented actor models [22, 141, 142]. Nonetheless, despite differences in their modeling assumptions [23], both of these approaches treat reciprocity as an observed network feature rather than a latent variable, as we do in our methods. However, integrating reciprocity into standard generative models presents a significant challenge. These models rely on the assumption of conditional independence between edges, implying that all interactions are considered independent given certain latent variables. This assumption is overly restrictive and limits the applicability of traditional methods in the analysis of real-world networks where reciprocity plays a crucial role. In this thesis, we addressed this issue and proposed alternative probabilistic methods that incorporate reciprocity into their mathematical framework, going beyond the conditional independence assumption. Specifically, our methods account for the pairwise dependencies between two directed edges connecting node pairs – a minimal relaxation that effectively capture reciprocity.

In Section 3.2.1 and Section 3.2.2, we introduced two different approaches to incorporate reciprocity into the framework of latent variable models. In particular, we combined both reciprocity and community structure within unique probabilistic methods for network analysis. While these patterns represent two separate mechanisms of network formation, their integration allows for more accurate and expressive generative models. Our first method, presented in Section 3.2.1 and referred to as CRep, is designed for the analysis of directed networks with nonnegative discrete weights [135]. In details, it employs Poisson distributions to model the conditional distributions $P(A_{ij} | A_{ji}, \Theta)$, where Θ encompasses the latent variables associated with reciprocity and community structure. Importantly, CRep expresses the network's likelihood using a pseudo-likelihood approximation, which involves a factorization over the specified conditionals rather than dealing with the unknown joint distributions $P(A_{ij}, A_{ji} | \Theta)$. In Section 3.2.2, we presented a different approach and introduced JointCRep, a latent variable model that explicitly describes the two-edge joint distributions, thus providing a closed-form expression for the network's likelihood [39]. It achieves this by employing bivariate Bernoulli distributions, making it particularly suitable for the analysis of binary directed networks. Despite the differences in their modeling assumptions, both of these models serve as foundational and valuable tools for the analysis of real-world data. In their respective papers, we showcased their properties, strengths, and limitations through extensive analyses of various synthetic and real-world datasets. Furthermore, we expanded upon these frameworks to explore other scenarios and applications, as presented in Section 3.2.3, Section 3.2.4 and Section 3.2.5.

Within Section 3.2.3, we introduced DynCRep, an extension of CRep tailored to accommodate

directed networks that evolve over time [136]. Fundamentally, DynCRep assumes that the evolution of interactions between two nodes in dynamic networks is influenced not only by the nodes' community memberships but also by their reciprocated interactions, whether occurring in the present or in the past. Notably, DynCRep was one of the first probabilistic models to take into account the role of reciprocity as an additional driver of network dynamics, proving its relevance in such contexts. In Section 3.2.4, instead, we took on the formalism of JointCRep to develop a generative model able to perform edge anomaly detection. This method, named CRAD, considers community membership and reciprocity as main mechanisms driving tie formation, and detects as anomalies those pairs of edges that deviate from this regular behavior. This model exclusively takes in input the adjacency matrix, making it particularly valuable in scenarios where additional information is unavailable. Conversely, in such cases, common anomaly detection models that rely on metadata face significant limitations in their applicability. Lastly, in Section 3.2.5, we outlined VIMuRe. This is a probabilistic model designed to estimate the unobserved network structure from multiply reported data. In this model, we integrated reciprocity by adopting the principles of CRep, reflecting the tendency of reporters to nominate the same individuals in both directions of a relationship [44]. Moreover, VIMuRe accommodates any number of reporters, allows the inclusion of weighted edges, and explicitly models individual biases in over- or under-reporting relationships through the incorporation of a reliability parameter. These characteristics make our model a powerful tool for the analysis of social networks, surpassing previous methods that rely on the union or intersection of reported data or exclusively focus on double-sampled data.

The methods discussed in this section benefit from efficient and scalable implementations, making use of either EM algorithms or VI techniques. Furthermore, they serve not only as tools for network inference, but also as benchmark models capable of generating synthetic data that align with the underlying assumptions of each model. Such benchmarks, which account for community structure and reciprocity to generate static, dynamic, anomalous, or multiply reported scenarios, were previously lacking in the field. With our models, we have equipped practitioners with the tools needed to test and compare their own methods, filling a significant gap in the field. In our experiments, we demonstrated the ability of these models in producing network samples that closely resemble the observed topological properties in the input data, including reciprocity, hierarchical structure, and degree distribution. Importantly, our methods outperformed standard generative models in this task, which are generally unable to reproduce the structural features of the network.

In our work, we loosened the common assumption of conditional independence by explicitly modeling either conditional or joint distributions. These methods not only introduced innovative perspectives for network modeling but also hold significant implications in inferential tasks. For instance, in link prediction problems, our models can now use conditional expected values to predict edge existence, alongside the conventional marginal expectations. Notably, our experiments consistently showcased better predictions when we used conditional expected values. This trend was also observed in network reconstruction tasks. In this context, our methods overcame the limitations of standard generative models, which struggle to recover reciprocated interactions due to the conditional independence assumption. This highlights the importance of effectively leveraging the extra information contained within the adjacency matrix to boost overall performance. Moreover, explicitly modeling pairwise dependencies increased results robustness, especially in scenarios characterized by high reciprocity, varying anomalies densities, or different structures of the affinity matrix over time. Furthermore, we illustrated how the formalism of JointCRep can be employed to predict the joint existence of mutual connections between pairs of nodes, providing principled and accurate outcomes in comparison to models that lack a specification for the joint distributions.

In addition to assessing our models' edge prediction abilities, we evaluated their performance in recovering the model parameters. We first investigated the community detection task, and we observed that JointCRep performed equally well as standard community detection algorithms, even with an additional parameter into the model. Moreover, its framework consistently achieved robust results, even in scenarios where anomalous edges were prevalent. On the other hand, CRep exhibited suboptimal performance in identifying communities in datasets characterized by high reciprocity. This limitation arises from the additive approach it uses to combine reciprocity and communities within the model. However, although this assumption may penalize the community detection task, it enables the estimation of the relative contributions of community and reciprocity in determining individual edges, a feature that JointCRep lacks due to its multiplicative combination. Subsequently, we showcased the ability of our models in capturing reciprocity, consistently outperforming other methods across various real-world networks. Shifting our focus to the extensions of our foundational models, we observed strong performance in anomaly detection with CRAD, highlighting how the integration of reciprocity can enhance performance compared to a model lacking this effect. Furthermore, VIMuRe demonstrated its proficiency in recovering the unobserved network from multiply reported data, improving standard deterministic aggregation approaches. Notably, VIMuRe also yielded more robust results in challenging scenarios where the number of unreliable reporters and reciprocity increased.

In this section, we extended the formalism of mixed-membership models to incorporate reciprocity, a relevant feature of real-world networks that influences their interactions. This was achieved by relaxing the conditional independence assumption and explicitly modeling the dependencies between pairs of nodes. Therefore, our models represent initial contributions to efficiently describe more complex scenarios, thereby improving our understanding of real-world data. Our initial focus was on single-layer networks, and potential future directions could involve extending these frameworks to data with additional information. As an example, we demonstrated this with DynCRep, where we incorporated time-varying interactions. Equally interesting would be the inclusion of node metadata and the exploration of how this extra information aligns with the reciprocity effect. Similarly, we could encompass multilayer networks, which raises the question of how properly integrate reciprocated edges from different layers. So far, we used a unique parameter to represent the reciprocity of the whole network. Nonetheless, in a multilayer context, it might be more appropriate to have distinct reciprocity parameters for each layer. One approach could be to model the interactions between the same pairs of nodes in all layers using a multivariate normal distribution, where the covariances represents reciprocity effects in those layers, extending ideas from single-layer networks [42]. Having a single parameter for reciprocity also limits our ability to capture individual tendencies of reciprocating relationships, which can vary depending on the roles of nodes in the network. Some social science models incorporate dyadic reciprocity parameters to describe the reciprocity of each pairwise interaction [129, 130]. However, these models serve different purposes than ours, primarily focusing on describing interactions – typically binary – and how they are influenced by hidden structures, rather than inferring these hidden patterns. Additionally, their applicability is limited to relatively small networks due to the computational cost associated with their sampling inferential techniques. Finally, a recent work [125] suggests potential for further improvements by extending SBMs to incorporate triadic closure, which accounts for dependencies involving triples rather than pairs. This work aligns more with the formalism of mixture models rather than combining mechanisms within a single model, making it slightly different from our proposed methods. Exploring the similarities and limitations of these approaches represents a promising avenue for future research, but it would require further steps to break conditional dependencies between edges to integrate these additional features within our frameworks.

4.3 Community detection and the analysis of higher-order data

Over the past few years, real-world data from diverse domains, including social and biological systems, have revealed interactions that go beyond pairwise connections, involving groups of nodes of various sizes. Hypergraphs provide a versatile and comprehensive framework for characterizing systems where such higher-order interactions are relevant. As explained in [Section 2.1](#), the inherent higher-order structure within these hypergraphs offers more realistic representations of real-world data, and their analysis can yield a deeper comprehension of the complex structure underlying these systems. This understanding can be achieved through the application of community detection algorithms, which are capable of identifying the mesoscale organization of real-world data. Several methods for detecting communities in hypergraphs have been proposed, including nonparametric methods with hypergraphons [8], flow-based algorithms [29, 54], and spectral clustering [5, 157]. Nevertheless, there are only a few probabilistic generative models that have been developed to rigorously define and identify the structural organization of hypergraphs [35, 110], making this area largely unexplored. In this thesis, we contributed to the advancement of these statistical inference techniques, expanding the toolkit available for the analysis of higher-order data.

In [Section 3.3.1](#) and [Section 3.3.2](#), we expounded two distinct probabilistic models designed to perform inference on hypergraphs while capturing their hidden organization. These models, named Hypergraph-MT [37] and Hy-MMSBM [134], posit the existence of a mixed-membership community structure as the main mechanism driving hyperedge formation. Such assumption was not explored in the analysis of hypergraphs before. In particular, our models are well-suited for analyzing nonnegative discrete weighted hyperedges, which are mathematically represented using Poisson distributions. The main difference between the two approaches lies in how they integrate latent variables into the model, leading to two different assumptions about data generation. Specifically, Hypergraph-MT expands upon Multitensor– the model presented in [Section 2.3.3](#) – and describes a hyperedge through the product of the memberships of all nodes belonging to it. To make this computation feasible, Hypergraph-MT assumes the existence of exclusively assortative community structures, which is a reasonable assumption in a variety of contexts. On the other hand, Hy-MMSBM relaxes the assortativity constraint and flexibly captures various community structures that were not tackled in the literature, such as disassortative and core-periphery, among others. To achieve this, it employs a bilinear form to link hyperedge probabilities and node community memberships. In addition to introducing these foundational models, we have also developed a new Python library, named `hypergraphx` [96], which provides a wide range of tools and algorithms for handling higher-order data. This computational effort, combined with comprehensive and user-friendly tutorials, enhances the usability of our methods. Furthermore, together with a few other recent and existing packages [6, 7, 93, 128], our contribution makes the analysis of real-world higher-order data more accessible, thus advancing our understanding of these complex systems.

The methods outlined in this section offer several advantages for the analysis of hypergraphs, and provide a good fit for their representations. We initially showcased their efficacy in community detection tasks across a diverse range of synthetic and real-world networks. In particular, we found that Hy-MMSBM consistently and accurately retrieved the planted communities in scenarios with varying hyperedges sizes. It also effectively captured assortative and disassortative community structures and correctly represented core-periphery configurations. Additionally, we observed that Hypergraph-MT detected communities that reliably depicted the information carried by hyperedges and exhibited robustness against the addition of noisy interactions. Subsequently, we investigated the ability of our methodologies in predicting missing hyperedges. Overall, our models outperformed

existing methods in predicting higher-order interactions of varying sizes. We also illustrated how our methods can leverage knowledge about large hyperedges to predict smaller ones, thus extracting valuable structural information from interactions of higher sizes. Lastly, it is important to emphasize that our models are highly efficient and scalable. They employ the EM algorithm to perform inference, and their numerical implementations make our methods computational and memory-efficient. This is not a trivial achievement, considering the increased information load associated with higher-order interactions. This property enhances the usability of our models, allowing for the study of real-world systems that were previously computationally challenging.

In this section, we introduced foundational methods that contributed to the development of statistical techniques for the analysis of hypergraphs. This field is relatively new and has received significant attention in recent years. As a result, several other approaches have been developed in the meantime [34, 91, 97, 139], however different from our framework. Focusing on our specific formalism, there exist numerous unexplored avenues for its further development. For instance, an interesting direction would be the incorporation of node attributes into our models, potentially enhancing their inference capabilities, as demonstrated in both single and multilayer networks. It may also be worthwhile to explore how our formalism could be adapted to accommodate edge metadata or directed hyperedges, moving in the direction of increasingly complex and multifaceted data representations. The exploration of temporal hypergraphs is another promising avenue, as they encode crucial information for understanding the chronological dynamics of interaction formation and evolution [30, 31, 57]. Other potential research directions involve the development of benchmark methods capable of generating synthetic data, facilitating a more comprehensive exploration of higher-order data. Recent examples include the use of Hy-MMSBM to create synthetic hypergraphs with overlapping communities and flexible structures [133], and the generation of random graph models with community structure and power-law distribution for both degrees and community sizes [83]. Given the relative novelty of this field, there is also a scarcity of theoretical studies on topics such as detectability thresholds and model identifiability, with only a few works dedicated to study the theoretical aspects of existing models [65, 66]. Furthermore, it would be worthwhile to investigate alternative formalism or probability distributions to gain a more comprehensive understanding of the strengths and limitations of statistical techniques in this field. Similarly, extending the concepts behind CRep and JointCRep to hypergraphs holds promise for developing more sophisticated methods that go beyond the conditional independence assumption, and incorporate sensible and general mechanisms that capture the complexity of real-world systems.

4.4 Conclusion

In this dissertation, we advanced the field of network inference by introducing statistical models that effectively capture the multifaceted complexities of real-world data. Specifically, we developed principled methods to account for the information embedded in attributed multilayer networks, to incorporate reciprocity by relaxing the conditional independence assumption, and to unveil the structural organization of higher-order data. These models, each tailored to tackle specific complexities and available through open-source implementations, confirmed the importance of analyzing more intricate representations to faithfully depict real-world data. Nonetheless, these models must continually evolve to effectively encompass the growing complexities of these systems. We hope that the methods discussed in this thesis can serve as a valuable foundation for the development of future probabilistic techniques, further enhancing our understanding of real-world systems.

Bibliography

- [1] Abbe, E. "Community detection and stochastic block models: recent developments". *The Journal of Machine Learning Research* 18.1 (2017), pages 6446–6531.
- [2] Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. "Mixed membership stochastic blockmodels". *Advances in Neural Information Processing Systems* 21 (2008).
- [3] Akaike, H. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19.6 (1974), pages 716–723.
- [4] Albert, R., Albert, I., and Nakarado, G. L. "Structural vulnerability of the North American power grid". *Physical Review E* 69.2 (2004), page 025103.
- [5] Angelini, M. C., Caltagirone, F., Krzakala, F., and Zdeborová, L. "Spectral detection on sparse hypergraphs". *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pages 66–73.
- [6] Antelmi, A., Cordasco, G., Kamiński, B., Prałat, P., Scarano, V., Spagnuolo, C., and Szufel, P. "Analyzing, exploring, and visualizing complex networks via hypergraphs using SimpleHypergraphs.jl". *Internet Mathematics* 1.1 (2020).
- [7] Badie-Modiri, A. and Kivelä, M. "Reticula: A temporal network and hypergraph analysis software package". *SoftwareX* 21 (2023), page 101301.
- [8] Balasubramanian, K. "Nonparametric modeling of higher-order interactions via hypergraphons". *The Journal of Machine Learning Research* 22.1 (2021), pages 6464–6498.
- [9] Ball, B., Karrer, B., and Newman, M. E. "Efficient and principled method for detecting communities in networks". *Physical Review E* 84.3 (2011), page 036103.
- [10] Banavar, J. R., Maritan, A., and Rinaldo, A. "Size and form in efficient transportation networks". *Nature* 399.6732 (1999), pages 130–132.
- [11] Barabási, A.-L. "Network science". *Cambridge University Press*, 2016.
- [12] Bascompte, J. "Disentangling the web of life". *Science* 325.5939 (2009), pages 416–419.
- [13] Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. "Networks beyond pairwise interactions: Structure and dynamics". *Physics Reports* 874 (2020), pages 1–92.
- [14] Battiston, F., Nicosia, V., and Latora, V. "Structural measures for multiplex networks". *Physical Review E* 89.3 (2014), page 032804.
- [15] Battiston, F. and Petri, G. "Higher-Order Systems". *Springer*, 2022.
- [16] Benson, A. R., Gleich, D. F., and Leskovec, J. "Higher-order organization of complex networks". *Science* 353.6295 (2016), pages 163–166.
- [17] Berge, C. "Graphs and hypergraphs". *North-Holland Pub. Co.*, 1973.
- [18] Bick, C., Gross, E., Harrington, H. A., and Schaub, M. T. "What are higher-order networks?" *SIAM Review* 65.3 (2023), pages 686–731.
- [19] Bishop, C. M. "Pattern recognition and machine learning". *Springer*, 2006.
- [20] Blei, D. M. "Build, compute, critique, repeat: Data analysis with latent variable models". *Annual Review of Statistics and Its Application* 1 (2014), pages 203–232.

- [21] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. "Variational inference: A review for statisticians". *Journal of the American Statistical Association* 112.518 (2017), pages 859–877.
- [22] Block, P. "Reciprocity, transitivity, and the mysterious three-cycle". *Social Networks* 40 (2015), pages 163–173.
- [23] Block, P., Stadtfeld, C., and Snijders, T. A. "Forms of dependence: Comparing SAOMs and ERGMs from basic principles". *Sociological Methods & Research* 48.1 (2019), pages 202–239.
- [24] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., and Zanin, M. "The structure and dynamics of multilayer networks". *Physics Reports* 544.1 (2014), pages 1–122.
- [25] Borgatti, S. P. and Everett, M. G. "Models of core/periphery structures". *Social Networks* 21.4 (2000), pages 375–395.
- [26] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. "Network analysis in the social sciences". *Science* 323.5916 (2009), pages 892–895.
- [27] Bullmore, E. and Sporns, O. "Complex brain networks: graph theoretical analysis of structural and functional systems". *Nature Reviews Neuroscience* 10.3 (2009), pages 186–198.
- [28] Cao, J., Jin, D., Yang, L., and Dang, J. "Incorporating network structure with node contents for community detection on large networks using deep learning". *Neurocomputing* 297 (2018), pages 71–81.
- [29] Carletti, T., Fanelli, D., and Lambiotte, R. "Random walks and community detection in hypergraphs". *Journal of Physics: Complexity* 2.1 (2021), page 015011.
- [30] Cencetti, G., Battiston, F., Lepri, B., and Karsai, M. "Temporal properties of higher-order interactions in social networks". *Scientific Reports* 11.1 (2021), page 7028.
- [31] Ceria, A. and Wang, H. "Temporal-topological properties of higher-order evolving networks". *Scientific Reports* 13.1 (2023), page 5885.
- [32] Chen, K. and Lei, J. "Network cross-validation for determining the number of communities in network data". *Journal of the American Statistical Association* 113.521 (2018), pages 241–251.
- [33] Chen, P.-Y. and Hero, A. O. "Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms". *IEEE Transactions on Signal and Information Processing over Networks* 3.3 (2017), pages 553–567.
- [34] Chodrow, P., Eikmeier, N., and Haddock, J. "Nonbacktracking spectral clustering of nonuniform hypergraphs". *SIAM Journal on Mathematics of Data Science* 5.2 (2023), pages 251–279.
- [35] Chodrow, P., Veldt, N., and Benson, A. R. "Generative hypergraph clustering: From block-models to modularity". *Science Advances* 7.28 (2021), eabh1303.
- [36] Côme, E. and Latouche, P. "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood". *Statistical Modelling* 15.6 (2015), pages 564–589.
- [37] Contisciani, M., Battiston, F., and De Bacco, C. "Inference of hyperedges and overlapping communities in hypergraphs". *Nature Communications* 13.1 (2022), page 7229.
- [38] Contisciani, M., Power, E. A., and De Bacco, C. "Community detection with node attributes in multilayer networks". *Scientific Reports* 10.1 (2020), page 15736.

- [39] Contisciani, M., Safdari, H., and De Bacco, C. "Community detection and reciprocity in networks by jointly modelling pairs of edges". *Journal of Complex Networks* 10.4 (2022), cnac034.
- [40] Coscia, M., Giannotti, F., and Pedreschi, D. "A classification for community discovery methods in complex networks". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4.5 (2011), pages 512–546.
- [41] Crucitti, P., Latora, V., and Marchiori, M. "A topological analysis of the Italian electric power grid". *Physica A: Statistical Mechanics and its Applications* 338.1-2 (2004), pages 92–97.
- [42] Dabbs, B., Adhikari, S., and Sweet, T. "Conditionally Independent Dyads (CID) network models: A latent variable approach to statistical social network analysis". *Social Networks* 63 (2020), pages 122–133.
- [43] Daudin, J.-J., Picard, F., and Robin, S. "A mixture model for random graphs". *Statistics and Computing* 18.2 (2008), pages 173–183.
- [44] De Bacco, C., Contisciani, M., Cardoso-Silva, J., Safdari, H., Lima Borges, G., Baptista, D., Sweet, T., Young, J.-G., Koster, J., Ross, C. T., et al. "Latent network models to account for noisy, multiply reported social network data". *Journal of the Royal Statistical Society Series A: Statistics in Society* 186.3 (2023), pages 355–375.
- [45] De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. "Community detection, link prediction, and layer interdependence in multilayer networks". *Physical Review E* 95.4 (2017), page 042317.
- [46] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". *Physical Review E* 84.6 (2011), page 066106.
- [47] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. "Inference and phase transitions in the detection of modules in sparse networks". *Physical Review Letters* 107.6 (2011), page 065701.
- [48] Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39.1 (1977), pages 1–22.
- [49] Dunne, J. A., Williams, R. J., and Martinez, N. D. "Food-web structure and network theory: the role of connectance and size". *Proceedings of the National Academy of Sciences* 99.20 (2002), pages 12917–12922.
- [50] Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. "Scale-free brain functional networks". *Physical Review Letters* 94.1 (2005), page 018102.
- [51] El Gheche, M., Chierchia, G., and Frossard, P. "OrthoNet: multilayer network data clustering". *IEEE Transactions on Signal and Information Processing over Networks* 6 (2020), pages 152–162.
- [52] Erdős, P. and Rényi, A. "On random graphs I". *Publicationes Mathematicae Debrecen* 6.290-297 (1959), page 18.
- [53] Erdős, P. and Rényi, A. "On the evolution of random graphs". *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* 5.1 (1960), pages 17–60.
- [54] Eriksson, A., Edler, D., Rojas, A., Domenico, M. de, and Rosvall, M. "How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs". *Communications Physics* 4.1 (2021), page 133.

- [55] Euler, L. "Solutio problematis ad geometriam situs pertinentis". *Commentarii academiae scientiarum Petropolitanae* (1741), pages 128–140.
- [56] Everett, B. "An introduction to latent variable models". Springer, 2013.
- [57] Failla, A., Citraro, S., and Rossetti, G. "Attributed Stream Hypergraphs: temporal modeling of node-attributed high-order interactions". *Applied Network Science* 8.1 (2023), pages 1–19.
- [58] Fajardo-Fontiveros, O., Guimerà, R., and Sales-Pardo, M. "Node metadata can produce predictability crossovers in network inference problems". *Physical Review X* 12.1 (2022), page 011010.
- [59] Faloutsos, M., Faloutsos, P., and Faloutsos, C. "On power-law relationships of the internet topology". *ACM SIGCOMM Computer Communication Review* 29.4 (1999), pages 251–262.
- [60] Fortunato, S. "Community detection in graphs". *Physics Reports* 486.3-5 (2010), pages 75–174.
- [61] Fortunato, S. and Hric, D. "Community detection in networks: A user guide". *Physics Reports* 659 (2016), pages 1–44.
- [62] Funke, T. and Becker, T. "Stochastic block models: A comparison of variants and inference methods". *PLOS ONE* 14.4 (2019), e0215296.
- [63] Garlaschelli, D. and Loffredo, M. I. "Patterns of link reciprocity in directed networks". *Physical Review Letters* 93.26 (2004), page 268701.
- [64] Ghasemian, A., Zhang, P., Clauset, A., Moore, C., and Peel, L. "Detectability thresholds and optimal algorithms for community structure in dynamic networks". *Physical Review X* 6.3 (2016), page 031005.
- [65] Ghoshdastidar, D. and Dukkipati, A. "Consistency of spectral partitioning of uniform hypergraphs under planted partition model". *Advances in Neural Information Processing Systems* 27 (2014).
- [66] Ghoshdastidar, D. and Dukkipati, A. "Consistency of spectral hypergraph partitioning under planted partition model". *The Annals of Statistics* 45.1 (2017), pages 289–315.
- [67] Gilbert, E. N. "Random graphs". *The Annals of Mathematical Statistics* 30.4 (1959), pages 1141–1144.
- [68] Girvan, M. and Newman, M. E. "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences* 99.12 (2002), pages 7821–7826.
- [69] Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. "A survey of statistical network models". *Foundations and Trends® in Machine Learning* 2.2 (2010), pages 129–233.
- [70] Gopalan, P. K., Gerrish, S., Freedman, M., Blei, D., and Mimno, D. "Scalable inference of overlapping communities". *Advances in Neural Information Processing Systems* 25 (2012).
- [71] Grilli, J., Barabás, G., Michalska-Smith, M. J., and Allesina, S. "Higher-order interactions stabilize dynamics in competitive network models". *Nature* 548.7666 (2017), pages 210–213.
- [72] Guimera, R., Mossa, S., Turtschi, A., and Amaral, L. N. "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles". *Proceedings of the National Academy of Sciences* 102.22 (2005), pages 7794–7799.
- [73] Hanley, J. A. and McNeil, B. J. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143.1 (1982), pages 29–36.

- [74] Higham, K., Contisciani, M., and De Bacco, C. "Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships". *Technological Forecasting and Social Change* 179 (2022), page 121628.
- [75] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. "Stochastic variational inference". *Journal of Machine Learning Research* (2013).
- [76] Hofman, J. M. and Wiggins, C. H. "Bayesian approach to network modularity". *Physical Review Letters* 100.25 (2008), page 258701.
- [77] Holland, P. W., Laskey, K. B., and Leinhardt, S. "Stochastic blockmodels: First steps". *Social Networks* 5.2 (1983), pages 109–137.
- [78] Holland, P. W. and Leinhardt, S. "An exponential family of probability distributions for directed graphs". *Journal of the American Statistical Association* 76.373 (1981), pages 33–50.
- [79] Holme, P. and Saramäki, J. "Temporal networks". *Physics Reports* 519.3 (2012), pages 97–125.
- [80] Hric, D., Peixoto, T. P., and Fortunato, S. "Network structure, metadata, and the prediction of missing nodes and annotations". *Physical Review X* 6.3 (2016), page 031038.
- [81] Jensen, J. L. W. V. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". *Acta Mathematica* 30.1 (1906), pages 175–193.
- [82] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. "An introduction to variational methods for graphical models". *Machine Learning* 37 (1999), pages 183–233.
- [83] Kamiński, B., Prałat, P., and Théberge, F. "Hypergraph Artificial Benchmark for Community Detection (h-ABCD)". *Journal of Complex Networks* 11.4 (2023), cnad028.
- [84] Karrer, B. and Newman, M. E. "Stochastic blockmodels and community structure in networks". *Physical Review E* 83.1 (2011), page 016107.
- [85] Keeling, M. J. and Eames, K. T. "Networks and epidemic models". *Journal of The Royal Society Interface* 2.4 (2005), pages 295–307.
- [86] Kim, D. I., Gopalan, P. K., Blei, D., and Sudderth, E. "Efficient online inference for bayesian nonparametric relational models". *Advances in Neural Information Processing Systems* 26 (2013).
- [87] Kim, M. and Leskovec, J. "Modeling Social Networks with Node Attributes using the Multiplicative Attribute Graph Model". *Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2011, 400–409.
- [88] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. "Multilayer networks". *Journal of Complex Networks* 2.3 (2014), pages 203–271.
- [89] Knoblauch, J., Jewson, J., and Damoulas, T. "An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference". *The Journal of Machine Learning Research* 23.1 (2022), pages 5789–5897.
- [90] Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.-K., Kishore, N., Hao, T., et al. "Network-based prediction of protein interactions". *Nature Communications* 10.1 (2019), page 1240.
- [91] Kritschgau, J., Kaiser, D., Rodriguez, O. A., Amburg, I., Bolkema, J., Grubb, T., Lan, F., Maleki, S., Chodrow, P., and Kay, B. "Community Detection in Hypergraphs via Mutual Information Maximization". *arXiv preprint arXiv:2308.04537* (2023).
- [92] Kullback, S. and Leibler, R. A. "On information and sufficiency". *The Annals of Mathematical Statistics* 22.1 (1951), pages 79–86.

- [93] Landry, N. W., Lucas, M., Iacopini, I., Petri, G., Schwarze, A., Patania, A., and Torres, L. “XGI: A Python package for higher-order interaction networks”. *Journal of Open Source Software* 8.85 (2023), page 5162.
- [94] Latora, V. and Marchiori, M. “Is the Boston subway a small-world network?” *Physica A: Statistical Mechanics and its Applications* 314.1-4 (2002), pages 109–113.
- [95] Lee, C. and Wilkinson, D. J. “A review of stochastic block models and extensions for graph clustering”. *Applied Network Science* 4.1 (2019), pages 1–50.
- [96] Lotito, Q. F., Contisciani, M., De Bacco, C., Di Gaetano, L., Gallo, L., Montresor, A., Musciotto, F., Ruggeri, N., and Battiston, F. “Hypergraphx: a library for higher-order network analysis”. *Journal of Complex Networks* 11.3 (2023), cnad019.
- [97] Lotito, Q. F., Musciotto, F., Montresor, A., and Battiston, F. “Hyperlink communities in higher-order networks”. *arXiv preprint arXiv:2303.01385* (2023).
- [98] Lusher, D., Koskinen, J., and Robins, G. “Exponential random graph models for social networks: Theory, methods, and applications”. *Cambridge University Press*, 2013.
- [99] Malliaros, F. D. and Vazirgiannis, M. “Clustering and community detection in directed networks: A survey”. *Physics Reports* 533.4 (2013), pages 95–142.
- [100] Matias, C. and Miele, V. “Statistical clustering of temporal networks through a dynamic stochastic block model”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.4 (2017), pages 1119–1141.
- [101] McLachlan, G. J. and Krishnan, T. “The EM algorithm and extensions”. *John Wiley & Sons*, 2007.
- [102] McPherson, M., Smith-Lovin, L., and Cook, J. M. “Birds of a feather: Homophily in social networks”. *Annual Review of Sociology* 27.1 (2001), pages 415–444.
- [103] Meng, X.-L. and Van Dyk, D. “The EM algorithm—an old folk-song sung to a fast new tune”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 59.3 (1997), pages 511–567.
- [104] Murphy, K. P. “Machine learning: a probabilistic perspective”. *MIT press*, 2012.
- [105] Newman, M. “Networks”. *Oxford University Press*, 2018.
- [106] Newman, M. E. “Communities, modules and large-scale structure in networks”. *Nature physics* 8.1 (2012), pages 25–31.
- [107] Newman, M. E. and Clauset, A. “Structure and inference in annotated networks”. *Nature Communications* 7.1 (2016), page 11863.
- [108] Newman, M. E. and Girvan, M. “Finding and evaluating community structure in networks”. *Physical Review E* 69.2 (2004), page 026113.
- [109] Newman, M. E. and Reinert, G. “Estimating the number of communities in a network”. *Physical Review Letters* 117.7 (2016), page 078301.
- [110] Ng, T. L. J. and Murphy, T. B. “Model-based clustering for random hypergraphs”. *Advances in Data Analysis and Classification* (2021), pages 1–33.
- [111] Nowicki, K. and Snijders, T. A. B. “Estimation and prediction for stochastic blockstructures”. *Journal of the American Statistical Association* 96.455 (2001), pages 1077–1087.
- [112] Opper, M. and Saad, D. “Advanced mean field methods: Theory and practice”. *MIT press*, 2001.

- [113] Ormerod, J. T. and Wand, M. P. “Explaining variational approximations”. *The American Statistician* 64.2 (2010), pages 140–153.
- [114] Otte, E. and Rousseau, R. “Social network analysis: a powerful strategy, also for the information sciences”. *Journal of Information Science* 28.6 (2002), pages 441–453.
- [115] Pagani, G. A. and Aiello, M. “The power grid as a complex network: a survey”. *Physica A: Statistical Mechanics and its Applications* 392.11 (2013), pages 2688–2700.
- [116] Papadopoulos, A., Pallis, G., and Dikaiakos, M. D. “Weighted clustering of attributed multi-graphs”. *Computing* 99 (2017), pages 813–840.
- [117] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. “Epidemic processes in complex networks”. *Reviews of Modern Physics* 87.3 (2015), page 925.
- [118] Peel, L., Larremore, D. B., and Clauset, A. “The ground truth about metadata and community detection in networks”. *Science Advances* 3.5 (2017), e1602548.
- [119] Peel, L., Peixoto, T. P., and De Domenico, M. “Statistical inference links data and theory in network science”. *Nature Communications* 13.1 (2022), page 6794.
- [120] Peixoto, T. P. “Entropy of stochastic blockmodel ensembles”. *Physical Review E* 85.5 (2012), page 056122.
- [121] Peixoto, T. P. “Parsimonious module inference in large networks”. *Physical Review Letters* 110.14 (2013), page 148701.
- [122] Peixoto, T. P. “Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models”. *Physical Review E* 89.1 (2014), page 012804.
- [123] Peixoto, T. P. “Hierarchical block structures and high-resolution model selection in large networks”. *Physical Review X* 4.1 (2014), page 011047.
- [124] Peixoto, T. P. “Bayesian stochastic blockmodeling”. *Advances in Network Clustering and Blockmodeling* (2019), pages 289–332.
- [125] Peixoto, T. P. “Disentangling homophily, community structure, and triadic closure in networks”. *Physical Review X* 12.1 (2022), page 011004.
- [126] Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. J., and Vaccarino, F. “Homological scaffolds of brain functional networks”. *Journal of The Royal Society Interface* 11.101 (2014), page 20140873.
- [127] Pizzuti, C. and Socievole, A. “A differential evolution-based approach for community detection in multilayer networks with attributes”. *Database and Expert Systems Applications: 31st International Conference, DEXA 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings, Part I* 31. Springer. 2020, pages 250–265.
- [128] Praggastis, B., Arendt, D., Joslyn, C., Purvine, E., Aksoy, S., and Monson, K. “HyperNetX”. *Pacific Northwest National Laboratory*. Available from: <https://github.com/pnrl/HyperNetX> (2019).
- [129] Redhead, D., Maliti, E., Andrews, J. B., and Borgerhoff Mulder, M. “The interdependence of relational and material wealth inequality in Pemba, Zanzibar”. *Philosophical Transactions of the Royal Society B* 378.1883 (2023), page 20220288.
- [130] Redhead, D., McElreath, R., and Ross, C. T. “Reliable network inference from unreliable data: A tutorial on latent network modeling using STRAND.” *Psychological Methods* (2023).
- [131] Redner, S. “How popular is your paper? An empirical study of the citation distribution”. *The European Physical Journal B-Condensed Matter and Complex Systems* 4.2 (1998), pages 131–134.

- [132] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. “An introduction to exponential random graph (p^*) models for social networks”. *Social Networks* 29.2 (2007), pages 173–191.
- [133] Ruggeri, N., Battiston, F., and De Bacco, C. “A principled, flexible and efficient framework for hypergraph benchmarking”. *arXiv preprint arXiv:2212.08593* (2022).
- [134] Ruggeri, N., Contisciani, M., Battiston, F., and De Bacco, C. “Community detection in large hypergraphs”. *Science Advances* 9.28 (2023), eadg9159.
- [135] Safdari, H., Contisciani, M., and De Bacco, C. “Generative model for reciprocity and community detection in networks”. *Physical Review Research* 3.2 (2021), page 023209.
- [136] Safdari, H., Contisciani, M., and De Bacco, C. “Reciprocity, community detection, and link prediction in dynamic networks”. *Journal of Physics: Complexity* 3.1 (2022), page 015010.
- [137] Safdari, H., Contisciani, M., and De Bacco, C. “Anomaly, reciprocity, and community detection in networks”. *Physical Review Research* 5.3 (2023), page 033084.
- [138] Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F., and Sanchez, A. “High-order interactions distort the functional landscape of microbial consortia”. *PLOS Biology* 17.12 (2019), e3000550.
- [139] Schlag, S., Heuer, T., Gottesbüren, L., Akhremtsev, Y., Schulz, C., and Sanders, P. “High-quality hypergraph partitioning”. *ACM Journal of Experimental Algorithmics* 27 (2023), pages 1–39.
- [140] Schwarz, G. “Estimating the dimension of a model”. *The Annals of Statistics* (1978), pages 461–464.
- [141] Snijders, T. A. “Stochastic actor-oriented models for network change”. *Journal of Mathematical Sociology* 21.1-2 (1996), pages 149–172.
- [142] Snijders, T. A. “The statistical evaluation of social network dynamics”. *Sociological Methodology* 31.1 (2001), pages 361–395.
- [143] Snijders, T. A. and Nowicki, K. “Estimation and prediction for stochastic blockmodels for graphs with latent block structure”. *Journal of Classification* 14.1 (1997), pages 75–100.
- [144] Solá, L., Romance, M., Criado, R., Flores, J., Amo, A. García del, and Boccaletti, S. “Eigenvector centrality of nodes in multiplex networks”. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23.3 (2013).
- [145] Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M., and Mucha, P. J. “Stochastic block models with multiple continuous attributes”. *Applied Network Science* 4.1 (2019), pages 1–22.
- [146] Torres, L., Blevins, A. S., Bassett, D., and Eliassi-Rad, T. “The why, how, and when of representations for complex systems”. *SIAM Review* 63.3 (2021), pages 435–485.
- [147] Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. “Modeling of protein interaction networks”. *Complexus* 1.1 (2003), pages 38–44.
- [148] Vespignani, A. “Modelling dynamical processes in complex socio-technical systems”. *Nature Physics* 8.1 (2012), pages 32–39.
- [149] Von Luxburg, U. “A tutorial on spectral clustering”. *Statistics and Computing* 17 (2007), pages 395–416.
- [150] Wainwright, M. J., Jordan, M. I., et al. “Graphical models, exponential families, and variational inference”. *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pages 1–305.
- [151] Wasserman, S. and Faust, K. “Social network analysis: Methods and applications”. *Cambridge University Press*, 1994.

- [152] Xu, S., Zhen, Y., and Wang, J. "Covariate-assisted community detection in multi-layer networks". *Journal of Business & Economic Statistics* 41.3 (2023), pages 915–926.
- [153] Yang, J., McAuley, J., and Leskovec, J. "Community detection in networks with node attributes". *2013 IEEE 13th International Conference on Data Mining*. IEEE. 2013, pages 1151–1156.
- [154] Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. "Detecting communities and their evolutions in dynamic social networks—a Bayesian approach". *Machine Learning* 82 (2011), pages 157–189.
- [155] Yook, S.-H., Jeong, H., and Barabási, A.-L. "Modeling the Internet's large-scale topology". *Proceedings of the National Academy of Sciences* 99.21 (2002), pages 13382–13386.
- [156] Zegura, E. W., Calvert, K. L., and Bhattacharjee, S. "How to model an internetwork". *Proceedings of IEEE INFOCOM'96. Conference on Computer Communications*. Volume 2. IEEE. 1996, pages 594–602.
- [157] Zhou, D., Huang, J., and Schölkopf, B. "Learning with hypergraphs: Clustering, classification, and embedding". *Advances in Neural Information Processing Systems* 19 (2006).

A Appendix

A.1 Published papers

In what follows, we present the ten peer-reviewed publications that are referenced within the dissertation.



OPEN

Community detection with node attributes in multilayer networks

Martina Contisciani¹, Eleanor A. Power² & Caterina De Bacco¹✉

Community detection in networks is commonly performed using information about interactions between nodes. Recent advances have been made to incorporate multiple types of interactions, thus generalizing standard methods to multilayer networks. Often, though, one can access additional information regarding individual nodes, attributes, or covariates. A relevant question is thus how to properly incorporate this extra information in such frameworks. Here we develop a method that incorporates both the topology of interactions and node attributes to extract communities in multilayer networks. We propose a principled probabilistic method that does not assume any *a priori* correlation structure between attributes and communities but rather infers this from data. This leads to an efficient algorithmic implementation that exploits the sparsity of the dataset and can be used to perform several inference tasks; we provide an open-source implementation of the code online. We demonstrate our method on both synthetic and real-world data and compare performance with methods that do not use any attribute information. We find that including node information helps in predicting missing links or attributes. It also leads to more interpretable community structures and allows the quantification of the impact of the node attributes given in input.

Community detection is a fundamental task when investigating network data. Its goal is to cluster nodes into communities and thus find large-scale patterns hidden behind interactions between many individual elements.

The range of application of this problem spans several disciplines. For instance, community detection has been used in sociology to analyze terrorist groups in online social networks¹; in finance to detect fraud events in telecommunication networks²; in engineering to refactor software packages in complex software networks³; and in biology to investigate lung cancer⁴ and to explore epidemic spreading processes⁵. In recent years, the variety of fields interested in this topic has broadened and the availability of rich datasets is increasing accordingly. However, most research approaches use only the information about interactions among nodes, in other words the network topology structure. This information can be complex and rich, as is the case for multilayer networks where one observes different types of interactions. For instance, in social networks, interactions could entail exchanging goods, socializing, giving advice, or requesting assistance. Most network datasets, however, contain additional information about individuals, attributes which describe their features, for instance their religion, age, or ethnicity. Node attributes are often neglected *a priori* by state-of-the-art community detection methods, in particular for multilayer networks. They are instead commonly used *a posteriori*, acting as candidates for “ground-truth” for real-world networks to measure the quality of the inferred partition^{6,7}, a practice that can also lead to incorrect scientific conclusions⁸. It is thus a fundamental question how to incorporate node attributes into community detection in a principled way. This is a challenging task because one has to combine two types of information⁹, while evaluating the extent to which topological and attribute information contribute to the network’s partition¹⁰.

To tackle these questions, we develop MTCOV, a mathematically rigorous and flexible model to address this problem for the general case of multilayer networks, i.e., in the presence of different types of interactions. The novelty of this model relies on a principled combination of the multilayer structure together with node information to perform community detection. To the best of our knowledge, MTCOV is the first overlapping community detection method proposed for multilayer networks with node attributes. The model leverages two sources of information, the topological network structure and node covariates (or attributes), to partition nodes into communities. It is flexible as it can be applied to a variety of network datasets, whether directed, weighted, or multilayer, and it outputs overlapping communities, i.e., nodes can belong to multiple groups simultaneously. In addition, the model does not assume any *a priori* correlation structure between the attributes and the

¹Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany. ²Department of Methodology, London School of Economics and Political Science, London WC2A 2AE, UK. ✉email: caterina.debacco@tuebingen.mpg.de

communities. On the contrary, the contribution of the attribute information is quantitatively given as an output of the algorithm by fitting the observed data. The magnitude of this contribution can vary based on the dataset. Even if this is not very high (for instance if the attributes are noisy or sparse) the model is nevertheless able to use this extra information to improve performance. At the same time, if incorporating attribute information hurts inference tasks, the model will downweigh this contribution and instead use mostly the topological network structure.

Our method allows domain experts to investigate particular attributes and select relevant community partitions based on what type of node information they are interested in studying. In fact, by choosing the input data, we can drive the algorithm to select for communities that are more relevant to the attribute under study. If the attribute hurts performance and is consequently downweighted by the algorithm, this can be used as a signal that the attribute might not correlate well with any partition, given the remaining topological information available, and thus inform the expert accordingly.

We study MTCOV on synthetic multilayer networks, a variety of single-layer node-attributed real networks and several real multilayer networks of social support interactions in two Indian villages. We measure performance based on prediction tasks and overlap with ground-truth (when this is known). For single-layer networks, we compare the performance of MTCOV to state-of-the-art community detection algorithms with node attributes; for multilayer networks, we test against a state-of-the-art algorithm that does not use any node attribute information and measure the extent to which knowing both types of information helps inference. We find that MTCOV performs well in predicting missing links and attributes. It also leads to more interpretable community structures and allows the quantification of the impact of the node attributes given as input.

To summarize, we present MTCOV, a new method that incorporates both the topology of interactions and node attributes to extract communities in multilayer networks. It is flexible, efficient and it has the property of quantitatively estimating the contributions of the two sources of information. It helps domain experts to investigate particular attributes and to better interpret the resulting communities. Moreover, by including relevant node attributes, it boosts performance in terms of edge prediction.

Related work. Several methods have been proposed to study community detection in networks¹¹. In particular, we are interested in those valid for multilayer networks¹². These generalize single-layer networks in that they can model different types of interactions and thus incorporate extra information that is increasingly available. Among these, we focus on generative models for multilayer networks^{13–19}, which are based on probabilistic modeling like Bayesian inference or maximum likelihood optimization. These are flexible and powerful in terms of allowing multiple inference tasks, injecting domain knowledge into the theoretical framework, and being computationally efficient. However, the majority of these methods do not consider node attributes as input along with the network information. In fact, the few methods developed for community detection in multilayer networks with node attributes are based on first aggregating the multilayer network into a single layer, either by combining directly the adjacency matrices of each layer²⁰ or by using similarity matrices derived from them along with the node attributes^{21,22}. In the context of data mining, a similar problem can be framed for learning low dimensional representations of heterogeneous data with both content and linkage structure (what we call attributes and edges). This is tackled by using embeddings extracted via deep architectures²³, which is rather different than our approach based on statistical inference. Our problem bears some common ground with the one studied by Sachan et al.²⁴ for extracting communities in online social networks, where users gather based on common interests; they adopt a Bayesian approach, but with a rather different goal of associating different types of edges to topics of interest. A related but different problem is that of performing community detection with node attributes on multiple independent networks^{25,26}; this differs with modeling a single multilayer network in that it assumes that covariates influence in the same way all the nodes in a network but in a different way the various networks in the ensemble. For single-layer networks, there has been more extensive work recently on incorporating extra information on nodes^{9,25,27–34}. Among those adopting probabilistic modeling, some incorporate covariate information into the prior information of the latent membership parameters^{25,35,36}, while others include covariates in an additive way along with the latent parameters^{37,38}, so that covariates influences the probability of interactions independently of the latent membership.

These works show the impact of adding nodes attributes in community detection *a priori* into the models to uncover meaningful patterns. One might then be tempted to adopt such methods also in multilayer networks by collapsing the topological structure into a suitable single network that can then be given in input to these single-layer and node-attributed methods as done by Gheche et al.²⁰. However, collapsing a multilayer network often leads to important loss of information, and one needs to be careful in determining when this collapse is appropriate and how it should be implemented, as shown for community detection methods without attribute information^{39,40}. Thus the need of a method that not only incorporates different types of edges but also node attributes.

Results

We test MTCOV's ability to detect communities in multilayer networks with node attributes by considering both synthetic and real-world datasets. We compare against MULTITENSOR¹³, an algorithm similar to ours but that does not include node attributes. We also test MTCOV's performance on single-layer networks, as the mathematical framework behind MTCOV still applies. Given this potential use and the paucity of algorithms suitable for comparison for multilayer networks, such comparisons assess the general utility of MTCOV.

Multilayer synthetic networks with ground-truth. To illustrate the flexibility and the robustness of our method, we generate multilayer synthetic networks with different kinds of structures in the various layers adapting the protocol described in De Bacco et al.¹³ to accommodate node attributes. We generate attributes as

Method	G ₁				G ₂				G ₃			
	F1-score	Jaccard	CS	L ₁	F1-score	Jaccard	CS	L ₁	F1-score	Jaccard	CS	L ₁
MULTITENSOR	0.512 ± 0.006	0.344 ± 0.006	0.585 ± 0.005	0.492 ± 0.004	0.514 ± 0.006	0.346 ± 0.06	0.614 ± 0.005	0.490 ± 0.005	0.999 ± 0.001	0.998 ± 0.001	0.991 ± 0.001	0.063 ± 0.002
MTCOV_0.3	0.7 ± 0.2	0.5 ± 0.2	0.7 ± 0.1	0.4 ± 0.1	0.8 ± 0.2	0.7 ± 0.2	0.8 ± 0.1	0.3 ± 0.2	0.995 ± 0.002	0.990 ± 0.004	0.984 ± 0.002	0.080 ± 0.004
MTCOV_0.5	0.6 ± 0.1	0.5 ± 0.2	0.7 ± 0.1	0.4 ± 0.1	0.992 ± 0.005	0.985 ± 0.009	0.986 ± 0.004	0.064 ± 0.004	0.996 ± 0.002	0.992 ± 0.004	0.985 ± 0.002	0.079 ± 0.004
MTCOV_0.7	0.988 ± 0.002	0.976 ± 0.004	0.977 ± 0.002	0.079 ± 0.003	1. ± 0.	1.000 ± 0.001	0.991 ± 0.001	0.062 ± 0.002	0.994 ± 0.002	0.988 ± 0.004	0.982 ± 0.001	0.087 ± 0.002
MTCOV_0.9	0.958 ± 0.003	0.920 ± 0.005	0.977 ± 0.001	0.050 ± 0.002	0.992 ± 0.002	0.984 ± 0.004	0.988 ± 0.001	0.050 ± 0.002	0.976 ± 0.003	0.952 ± 0.006	0.982 ± 0.002	0.051 ± 0.003

Table 1. Performance of algorithms MULTITENSOR and MTCOV on synthetic multilayer networks with attributes. We use different matches (one per row, e.g., MTCOV_0.3 denotes a match of 0.3, this is also the value we use to fix γ) between attributes and planted communities on synthetic directed multilayer networks. Results are averages and standard deviations over 10 networks samples for each network type G_m , $m = 1, 2, 3$; we take the average performance over the incoming and outgoing memberships, i.e., the matrices U and V , and the best performances are in boldface. Networks are generated with stochastic block model with $C = 2$, $N = 1000$ and average degree $k = 4$. Denote W^a, W^d, W^{cp} and W^{bd} , the affinity matrices of the assortative, disassortative, core-periphery and the biased directed layers respectively. Then, their entries are $w_{11}^a = w_{22}^a = w_{12}^d = w_{21}^d = w_{11}^{cp} = w_{12}^{bd} = \frac{kC}{N}$, $w_{12}^a = w_{21}^a = w_{11}^d = w_{22}^d = w_{12}^{cp} = w_{21}^{bd} = w_{11}^{bd} = w_{22}^{bd} = 0.1 \times \frac{kC}{N}$ and $w_{22}^{cp} = w_{12}^{bd} = 0.03 \times \frac{kC}{N}$. The F1-score, Jaccard, CS and L₁ are performance metrics as defined in the “Methods” section.

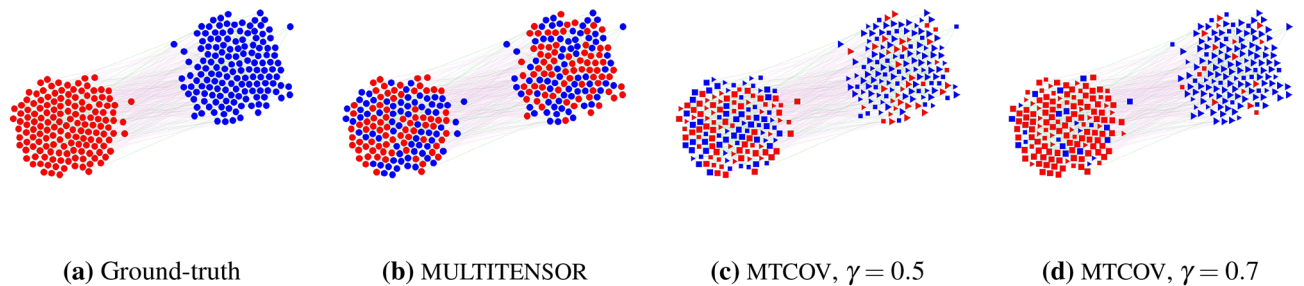


Figure 1. Partition of a synthetic multilayer network with attributes. We generated synthetic directed multilayer networks using a stochastic block model, that aligns with G_1 . To illustrate, here we do the equivalent task on a smaller network of size $N = 299$, $C = 2$ communities of equal-size unimixed group membership and $L = 2$ layers, of which one is assortative (green) and one disassortative (pink); (a) the ground-truth partition; (b–d) the communities inferred by three different methods: (b) MULTITENSOR, an algorithm without attributes, (c) MTCOV using the network structure and the attributes with the same proportion, i. e. $\gamma = 0.5$ and (d) MTCOV using mostly the attribute structure, i.e. $\gamma = 0.7$. Colors denote the inferred partition; the attributes in (c) and (d) are generated by matching them with true community assignments for the 50% and 70% of the nodes respectively, and chosen uniformly at random from the non-matching values; square and triangle denote the synthetic dummy attribute (squares are matched with the red group, triangles with the blue) and the size of the node shows the nodes matched with the true community (bigger means deterministic match, smaller means uniform at random match). We use the matrix U for the membership.

done in Newman and Clauset²⁷: we match them with planted communities in increasing ratios varying from 0.3 to 0.9; these values correspond also to the γ parameters that we fix for MTCOV. Specifically, we generate three types of directed networks using a stochastic block model⁴¹, all with $C = 2$ communities of equal-size unimixed group membership and $N = 1000$ nodes, but with different numbers and kinds of layers, similar to De Bacco et al.¹³. The first network (G_1) has $L = 2$ layers, one assortative (W^α has higher diagonal entries) and one disassortative (W^α has higher off-diagonal entries); the second (G_2) has $L = 4$ layers, two assortative and two disassortative and the third (G_3) has $L = 4$ layers, one assortative, one disassortative, one core-periphery (W^α has higher diagonal entries but one of the two is bigger than the other) and one with biased directed structure (W^α has higher off-diagonal entries but one of the two is bigger than the other). We generate ten independent samples of each of these types of networks and use all the evaluation metrics described in the “Methods” section in the presence of ground-truth. We use the membership inferred by the algorithms using the best maximum likelihood fixed point over 10 runs with different random initial conditions. As shown in Table 1, MTCOV performs significantly better than MULTITENSOR on the first and second network. This suggests that incorporating attribute information can significantly boost inference, with an increasing benefit for a smaller number of layers. Figure 1 shows an example of this result. Notice that G_2 requires a smaller match ($\gamma = 0.5$) between attributes and communities than G_1 ($\gamma = 0.7$) to achieve similar performance. G_1 and G_2 have similar structure, but the second has twice as many layers. Thus, increasing the number of layers may require less contribution from the extra information of the attributes, a possible advantage for multilayer networks. This intuition is reinforced by noticing not only that the best performance is achieved for $\gamma < 0.9$, but also that both the algorithms perform very well in the third network, regardless of the value of the match between attributes and communities. Con-

Village	Year	Nodes	Edges	$\langle k \rangle$	Caste	Religion	Age	Gender
Aḷakāpuram	2013	419	4,161	20	14	3	11	2
	2017	441	5,578	25	13	3	12	2
Tenpaṭṭi	2013	362	3,374	19	11	2	11	2
	2017	346	3,806	22	9	2	12	2

Table 2. Network summary statistics for the four social support networks of Indian villages. Each has the same set of 6 layers and Edges are the total over them; $\langle k \rangle$ is the average degree per node on the whole multilayer network. The columns Caste, Religion, Age and Gender are the number of different categories observed in each network for their respective attribute.

trary to G_2 , G_3 has a different structure in each of the 4 layers. This diversity can be even more beneficial than having more but correlated layers (as in G_1 vs. G_2). These synthetic tests demonstrate the impact of leveraging both node attributes and topological information: when topological structure is not very informative (as in G_1 with only two layers), adding node attributes can significantly help in recovering the communities. In contrast, when topological information is more complex (as in G_3 where all layers are different), properly combining the different layers' structures can compensate for a limited access to extra information on nodes. Overall, this shows the need for methods suited for exploiting various sources of information and the complexity behind multilayer networks.

Multilayer social support network of rural Indian villages. We demonstrate our model beyond synthetic data by applying it to social support networks of two villages in Tamil Nadu, India, which we call by the pseudonyms “Tenpaṭṭi” (Ten) and “Aḷakāpuram” (Ala)^{42–44}. Data were collected in the form of surveys where adult residents were asked to nominate those individuals who provided them with various types of support, including running errands, giving advice, and lending cash or other household items. These were collected in two rounds, one in 2013 and the other in 2017. Each type of support corresponds to a layer in the network; we consider only those layers present in both rounds, for a total of $L = 6$ layers. After pre-processing the data, by considering only those individuals who had at least one outgoing edge and removing self-loops, the resulting networks have the size reported in Table 2. In addition, several attributes were collected, which include information about age, religion, caste, and education level. Ethnographic observation in these villages⁴² and previous analyses^{43,44} suggest that social relations are strongly structured by religious and caste identity, with these divisions shaping where people live, who they marry, and who they choose to associate with. In other words, they suggest a dependence between the attributes Religion and Caste and the mechanisms driving edge formation in these social support networks. Motivated by these insights, here we consider the attributes Caste and Religion and add them into the model. In addition, we test the importance of variables that we expect to be less informative, such as gender and age. The latter, being continuous, is also an example of a non-categorical variable. Provided it has a finite range, as it is the case for age, we can encode it into categorical by binning its values. Here we use equal bins of size 5 years.

Without assuming *a priori* any ground-truth, we measure performance using the AUC and accuracy as explained in the “Methods” section. We compare with MULTITENSOR to measure the extent to which adding the attributes helps predicting edges and attributes; in addition, in terms of accuracy values, we consider two baselines for further comparisons: (1) a uniform at random probability over the number of possible categories (RP); and (2) the maximum relative frequency of the attribute value appearing more often (MRF). We fix hyperparameters using 5-fold cross-validation along with grid-search procedure (see “Cross-validation tests and hyperparameter settings” section for more details). We obtain values of $\gamma \in [0.2, 0.9]$, signalling relevant correlations between attributes and communities. For details, see Supplementary Table S2. Empirically, we observe that when $\gamma > 0.5$ the algorithm achieves better performance in terms of link and attribute prediction by well balancing the log-likelihood of the attribute dimension and the one of the network structure.

For validation, we split the dataset into training/test sets uniformly at random as explained in the “Methods” section. Table 3 reports the average results over ten runs for each network, and shows that MTCOV is capable of leveraging two sources of information to improve both performance metrics. In fact, our algorithm systematically achieves the highest accuracy for attribute prediction and the highest AUC for edge prediction (boldface). While a good performance in attribute prediction is expected by design as we add this data into the model, the fact that it also boosts performance in terms of edge prediction is not granted *a priori*. Instead, it is a quantitative way to show that an attribute plays an important role in the system. It also demonstrates the potential of capturing correlations between two different sources of information, which can have relevant applications, in particular when missing information of one kind. Notice in particular the improvement in AUC when using caste compared to no attribute given (MULTITENSOR). The other attributes are less informative; in particular age has a performance similar to MULTITENSOR in edge prediction, signalling that it does not contribute to inform edge formation. Indeed, it has the smallest inferred γ (always < 0.5), which gives also worse accuracy performance than the baseline, signalling again that this attribute may not be correlated with the community structure. All these results show the flexibility of MTCOV in adapting based on the data given in input: if warranted, it is able to ignore those attributes that are not correlated with network division and instead find communities that are mainly based on the network structure. Next, we test how adding node attributes impacts robustness against unbalanced data, where the ratio of positive examples (existing edges) observed in the training is different than that in the test set. We denote the total probability of selecting an edge in the test as *tpe*

Attribute	Method	ACCURACY for attribute prediction				AUC for link prediction			
		Ala 2013	Ala 2017	Ten 2013	Ten 2017	Ala 2013	Ala 2017	Ten 2013	Ten 2017
	MULTITENSOR					0.771 ± 0.009	0.835 ± 0.006	0.758 ± 0.005	0.81 ± 0.01
Caste	RP	0.07	0.08	0.10	0.11				
	MRF	0.556 ± 0.009	0.57 ± 0.01	0.32 ± 0.01	0.33 ± 0.02				
	MTCOV	0.80 ± 0.05	0.77 ± 0.05	0.69 ± 0.09	0.74 ± 0.07	0.837 ± 0.009	0.858 ± 0.008	0.829 ± 0.006	0.82 ± 0.01
Religion	RP	0.33	0.33	0.50	0.50				
	MRF	0.837 ± 0.008	0.843 ± 0.006	0.696 ± 0.008	0.679 ± 0.008				
	MTCOV	0.96 ± 0.02	0.95 ± 0.03	0.76 ± 0.08	0.80 ± 0.05	0.813 ± 0.007	0.83 ± 0.01	0.81 ± 0.02	0.80 ± 0.01
Age	RP	0.09	0.08	0.09	0.08				
	MRF	0.135 ± 0.005	0.126 ± 0.007	0.126 ± 0.005	0.128 ± 0.008				
	MTCOV	0.11 ± 0.03	0.11 ± 0.02	0.13 ± 0.04	0.10 ± 0.03	0.80 ± 0.01	0.823 ± 0.008	0.783 ± 0.009	0.80 ± 0.01
Gender	RP	0.50	0.50	0.50	0.50				
	MRF	0.584 ± 0.009	0.58 ± 0.01	0.56 ± 0.01	0.55 ± 0.01				
	MTCOV	0.61 ± 0.05	0.65 ± 0.04	0.58 ± 0.08	0.71 ± 0.08	0.79 ± 0.02	0.831 ± 0.009	0.80 ± 0.01	0.81 ± 0.01

Table 3. Prediction performance on real multilayer networks with attributes. Results are averages and standard deviations over 10 independent trials of cross-validation with 80–20 splits selected uniformly at random (i.e., $tpe = 0.004$); the best performances are in boldface. Datasets are described in Table 2. RP is the performance of uniform random probability and MRF the one of the maximum relative frequency, see “Methods” section for details.

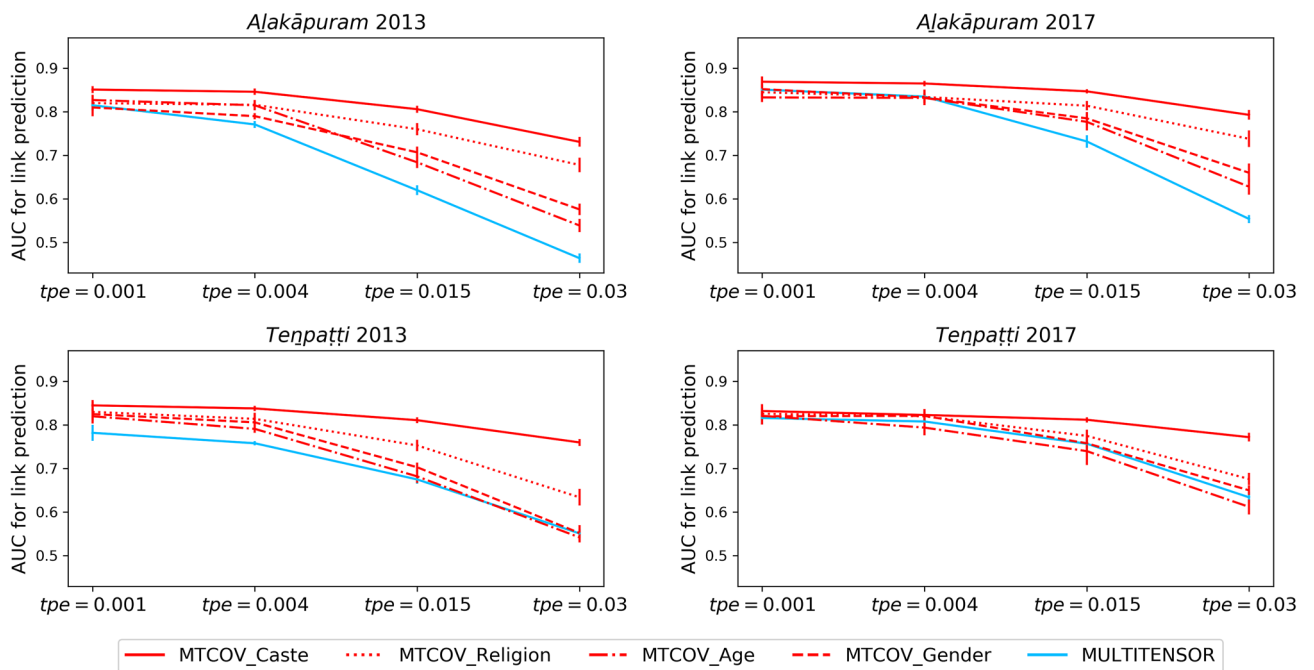


Figure 2. Probabilistic link prediction with biased edge sampling. Results are AUC values of MTCOV and MULTITENSOR on four social support networks in different held-out settings. Here tpe indicates the total probability of selecting one edge (positive example) in the test set. We consider Caste, Religion, Age and Gender attributes; results are averages and standard deviations over 10 independent runs.

and consider values $tpe \in \{0.001, 0.004, 0.015, 0.03\}$ denoting under-representation (0.001), equal (0.004), and over-representation (values 0.015 and 0.03) compared to the uniform at random selection (empirically we find $tpe = 0.004$). In these tests, we hold out 20% of the entries of A biasing their selection using the tpe values; in addition, we give as input the whole design matrix X (attributes) and measure link prediction performance. We observe that MTCOV is significantly more robust than the algorithm that does not use any attribute information, regardless of the value of γ . In fact, even though performance deteriorates as we decrease the number of positive examples in the training set (i.e., higher tpe), MTCOV is less impacted by this, as shown in Fig. 2 (results reported in Supplementary Table S3). Notice in particular performance discrepancies when using the attribute

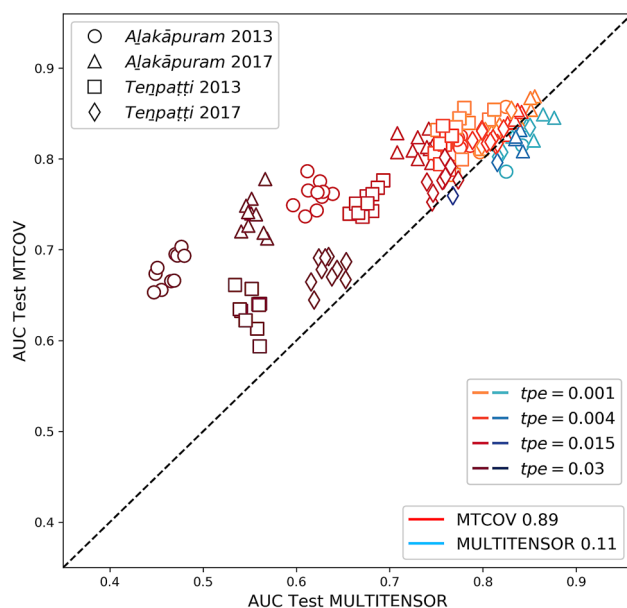


Figure 3. Trial-by-trial probabilistic link prediction with biased edge sampling. The values of AUC for MTCOV and MULTITENSOR are shown on the vertical axis and the horizontal axis respectively. The brightness represents the hardness of the settings in terms of biasing the edge sampling in the training. From bottom to top: $tpe = 0.03$ (hard, dark color), $tpe = 0.015$, $tpe = 0.004$ (random), $tpe = 0.001$ (easy, light color). Points above the diagonal, shown in shades of red, are trials where MTCOV is better performing than MULTITENSOR. The fractions for which each method is superior are shown in the plot legend. We use the attribute Religion.

Caste in the difficult regimes ($tpe \in \{0.015, 0.03\}$): MTCOV's performance deteriorates only a little, while using the other attributes or no attribute makes performance significantly worse, with AUC down to 0.6 from a value higher than 0.8 for easier regimes. Moreover, notice that attributes with the same scaling parameter value can give different prediction results, underlying the necessity to consider both the value of the estimated γ and the quality of the attribute to quantify its importance. This could explain why Caste provides always better results, given by the fact that its categories are more heterogeneous (i.e., more information) than Religion and Gender. The robustness of MTCOV is also confirmed by analyzing the performances on a trial-by-trial basis, each trial being a random sample of the held-out entries. As we show in Fig. 3, MTCOV better predicts links in 89% of the trials and never goes below the threshold of 0.5, the baseline random choice. These results demonstrate how adding another source of information helps when observing a limited amount of network edges.

Qualitative analysis of a social support network. To demonstrate our MTCOV model beyond prediction tasks and highlight its potential for interpretability, we show as an example its qualitative behavior on the real network of Ajakapuram in 2017 (see Table 2). Specifically, we compare the communities extracted by our algorithm and those inferred by MULTITENSOR. To ease comparison, we fix the same number of groups to $C = 4$ for both algorithms and measure how caste membership distributes across communities, and fix $\gamma = 0.8$ as obtained with cross-validation. Figure 4 shows the magnitude of each individual's inferred outgoing memberships u_i in each group. While the communities identified by MTCOV and MULTITENSOR show substantial similarities, MTCOV generally classifies castes more consistently into distinct communities, as we show in Figs. 4 and 5. To make a quantitative estimate of the different behaviors, we measure the entropy of the attribute inside each community $H_k = -\sum_{z=1}^Z f_z \log f_z / \log(Z)$, where f_z is the relative frequency of the z -th caste inside a group k , and the denominator is the entropy of a uniform distribution over the Z castes, our baseline for comparison. Values of H_k close to 1 denote a more uniform distribution of castes, whereas smaller values denote an unbalanced distribution with most of the people belonging to a few castes. We find that MTCOV has smaller entropies over the groups, with two groups having the smallest values, whereas MULTITENSOR has the highest, showing its tendency to cluster individuals of different castes into the same group. In addition, we observe that MTCOV has a more heterogeneous group size distribution which seems to be correlated with caste. Notably, the algorithms differ in how they place two caste groups that live in hamlets separated from the main village (the Hindu Yātavars and CSI Paraiyars). With MULTITENSOR, they are grouped together, while with MTCOV, the Hindu Yātavars are joined up into a community with Paḷḷars and Kulāḷars. While MULTITENSOR is clearly picking up the structural similarities of the two hamlets, this division makes little sense socially and culturally. In contrast, the way in which MTCOV defines a community which spans caste boundaries (MTCOV C1) aligns with ethnographic knowledge of the relations between these castes. Finally, we remark that there might be multiple meaningful community divisions in the network, and the fact that MTCOV's partition seems to better capture the distributions in the attribute caste does not mean that one algorithm is better than the other. In fact, there might be other hidden topological properties that MULTITENSOR's partition is picking up by being agnostic to

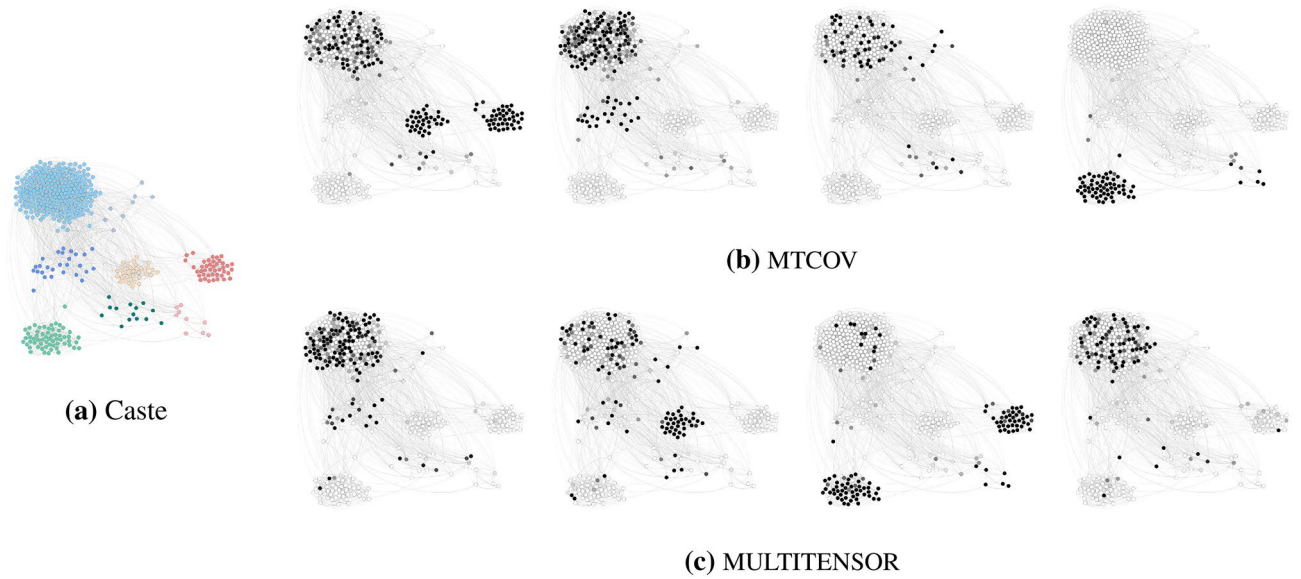


Figure 4. Attributes and inferred communities. Nodes of the social support network of Alakapuram in 2017 are colored by: (a) the attribute Caste (with colors as shown in Fig. 5); inferred communities by (b) MTCOV and (c) MULTITENSOR. Darker values in the grey scales indicate higher values of the entry of the membership vector u_i .

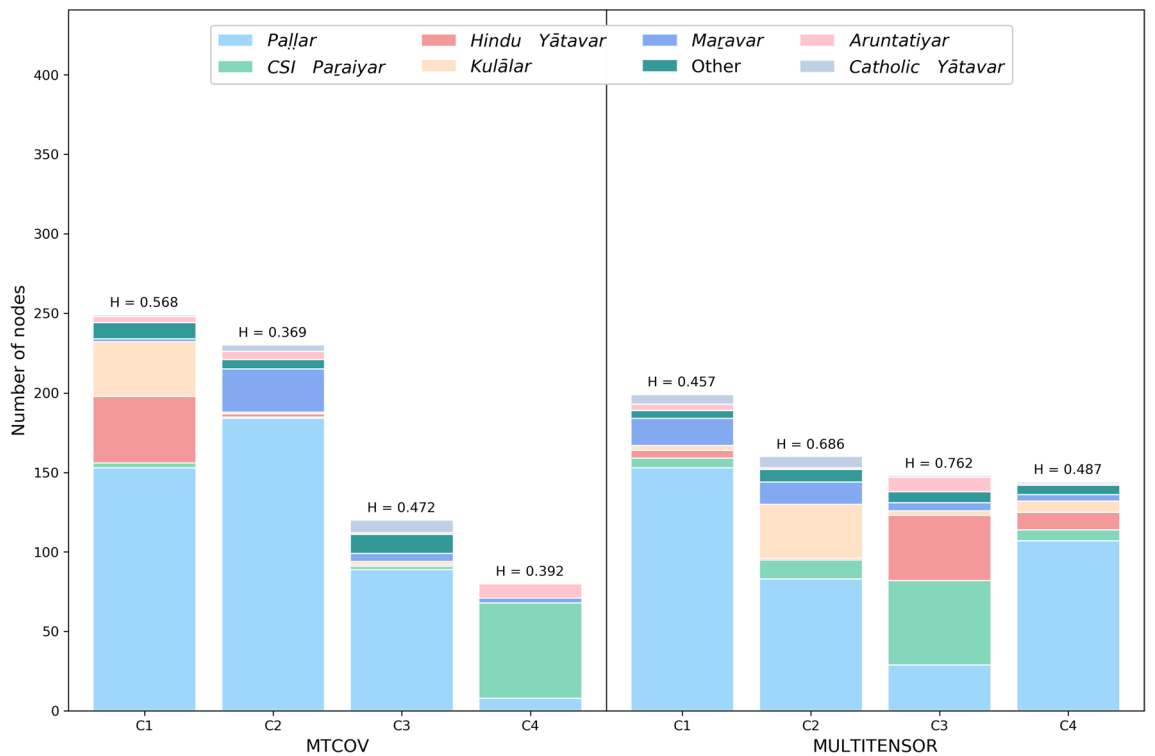


Figure 5. Partition of the attribute Caste inside each community detected by MTCOV and MULTITENSOR in the social support network of Alakapuram in 2017. The category Other contains small categories having less than five individuals. The label on top of each bar is the value of the entropy of the variable Caste inside the corresponding community. Note that nodes can have mixed membership, here we build a group k by adding to it all nodes i that have a non-zero k -th entry u_{ik} . The number of nodes is $N = 441$, corresponding to the maximum value of the y -axis plotted.

Method	F1-score			Jaccard similarity		
	facebook	football	polblogs	facebook	football	polblogs
MTCOV	0.5 ± 0.1	0.86 ± 0.03	0.8 ± 0.2	0.4 ± 0.1	0.82 ± 0.04	0.8 ± 0.2
NC	0.48 ± 0.08	0.82 ± 0.06	0.95 ± 0.09	0.36 ± 0.08	0.75 ± 0.08	0.9 ± 0.1
CESNA	0.46 ± 0.09	0.7 ± 0.0	0.6 ± 0.0	0.33 ± 0.08	0.6 ± 0.0	0.4 ± 0.0

Table 4. Performance of methods MTCOV, NC and CESNA on three datasets, according to two different measures used in the Eq. (16). The results are averages and standard deviations over ten independent runs and the best outcomes are bolded.

caste membership. The choice of which algorithm to use should be made based on the final goal of the application at hand.

Results on single-layer networks. Our model can be used for single-layer networks as well. For these we can compare against two state-of-the-art algorithms, both probabilistic generative models but different in their assumptions: CESNA⁹ which considers overlapping communities and posits two independent Bernoulli distributions for network edges and node attributes; and the model proposed by Newman and Clauset²⁷ (NC) for non-overlapping communities, a Bayesian approach where the priors on the community memberships depend on the node attributes. CESNA, similarly to our model, assumes conditional independence of the two likelihoods and introduces a regularization parameter between them; it uses block-coordinate ascent for parameters' estimation, while NC uses an EM algorithm for parameters' estimation, similarly to what we do here. We test MTCOV against them on both synthetic and real single-layer networks with node attributes, with and without ground-truth. We transform directed networks to undirected because both CESNA and NC do not distinguish for edge directionality. Results on synthetic data show that MTCOV and NC have similar performance in correctly classifying nodes in their ground-truth communities and both are better than CESNA; the main difference is that MTCOV is more stable and has less variance for high attribute correlation, in particular in the hard regime where classification is more difficult. We leave details in the Supplementary Section S4. For single-layer real networks, we use datasets with ground-truth candidates and node attributes: the ego-Facebook network (*facebook*)⁴⁵, a set of 21 networks built from connections between a person's friends where potential ground-truth are circles of friends hand-labeled by the ego herself; the American College football network (*football*)⁴⁶, a network of football teams playing against each other, where a ground-truth candidate is the conference to which each team belongs; and a network of political blogs (*polblogs*)⁴⁷ where potential ground-truth communities are divided by *left/liberal* and *right/conservative* political parties, see Supplementary Section S4 for details. For each network, we run a 5-fold cross-validation procedure combined with grid-search for fixing the hyperparameter γ (see "Cross-validation tests and hyperparameter settings" section for details; note that in this case we use the ground-truth value of C , hence γ is the only hyperparameter left to be tuned). For *facebook* we find that the average over the 21 networks is $\gamma = 0.15$, which signals a low correlation between the covariates and the communities, whereas for the *football* and *polblogs* networks we obtain much higher values of γ equal to 0.6 and 0.75 respectively. MTCOV has better performance in terms of F1-score and Jaccard similarity across the majority of datasets, as shown in Table 4. This is also supported by a trial-by-trial comparison shown in Fig. 6 for F1-score (similar results are obtained for Jaccard), where we find that MTCOV is more accurate in 59% and 90% of the cases than NC and CESNA, respectively.

Discussion

We present MTCOV, a generative model that performs overlapping community detection in multilayer networks with node attributes. We show its robustness in adapting to different scenarios, and its flexibility in exploiting the attributes that are more informative while ignoring those that are less correlated with the network communities. Our method is capable of estimating quantitatively the contribution given by the attributes and incorporating them to improve prediction performance both in terms of recovering missing attributes and in terms of link prediction. This allows domain experts to investigate particular attributes and select relevant community partitions based on what type of node information they are interested in investigating. There are valuable possible extensions of this work. One example is to incorporate modeling of more complex data types for the attributes, for instance combinations of discrete and continuous attributes, or other types of extra information, like time-varying network elements, whether the attributes, node, edges or combinations of these. From a technical point of view, when the topological and attribute datasets are very unbalanced in size, this might impact their relative likelihood weight and thus inference. One should then consider automating the process of rescaling them accordingly, as a pre-processing step to be incorporated into the model. Similarly, hyperparameter selection would benefit from an automatized routine when more than one performance metric is considered. The relations between attributes and communities could be transferred across networks to predict missing information when having access to similar but incomplete datasets. We show examples of these here, where we studied two snapshots of the same village networks across time. While we leave these questions for future work, we provide an open source version of the code.

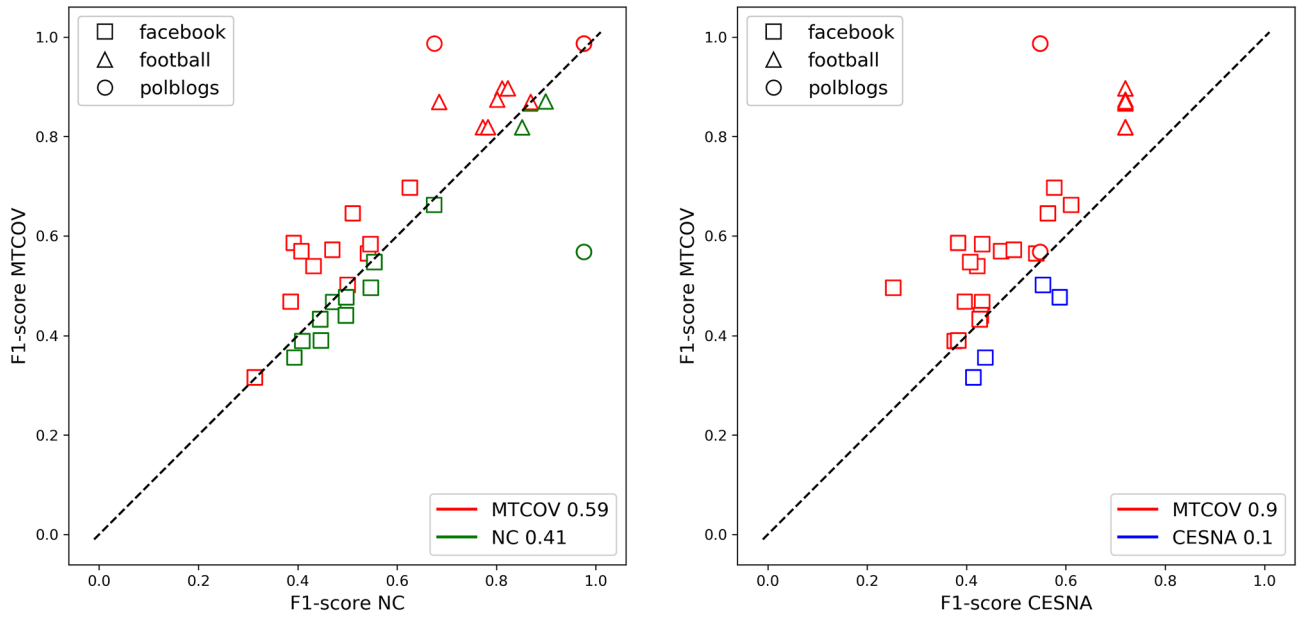


Figure 6. Trial-by-trial performance in F1-score. We compare MTCOV on the y-axis, with on the x-axis (left) NC and (right) CESNA. Markers denote the datasets: squares for *facebook*, triangles for *football* and circles for *polblogs*. Points above the diagonal, shown in red, are trials where MTCOV is more accurate than the other. The fractions for which each method is superior are shown in the plot legend.

Methods

We adapt recent ideas from the generative model behind MULTITENSOR¹³, a multilayer mixed-membership model based on a Poisson tensor factorization⁴⁸, to incorporate node attributes in a principled manner. It can take in input directed and undirected networks, allowing different topological structures in each layer, including arbitrarily mixtures of assortative, disassortative and core-periphery structures. We move beyond MULTITENSOR by incorporating node covariates via introducing a proper likelihood term that accounts for this extra information. We use the formalism of maximum likelihood estimation: we combine the structural and the node information into a global likelihood function and provide a highly scalable Expectation-Maximization algorithm for the estimation of parameters.

Model description and notation. Consider a multilayer network of N nodes and L layers. This is a set of graphs $G = \{G^{(\alpha)}(\mathcal{V}, \mathcal{E}^{(\alpha)})\}_{1 \leq \alpha \leq L}$ defined on a set \mathcal{V} of N vertices shared across $L \geq 1$ layers, and $\mathcal{E}^{(\alpha)}$ is the set of edges in the layer α . Each layer $\alpha \in \{1, \dots, L\}$ is a graph $G^{(\alpha)}(\mathcal{V}, \mathcal{E}^{(\alpha)})$ with adjacency matrix $A^{(\alpha)} = [a_{ij}^{(\alpha)}] \in \mathbb{R}^{N \times N}$, where $a_{ij}^{(\alpha)}$ is the number of edges of type α from i to j ; here we consider only positive discrete entries; for binary entries, $E = \sum_{i,j,\alpha} a_{ij}^{(\alpha)}$ is the total number of edges. Alternatively, we can consider a 3-way tensor A with dimensions $N \times N \times L$. In addition, for each node $i \in \mathcal{V}$ consider the vector of covariates $X_i \in \mathbb{R}^{1 \times K}$ (alternatively called also attributes or metadata), where K is the total number of attributes. Here, for simplicity we focus on the case of $K = 1$ and categorical covariates with Z different categories. However, we can easily generalize to more than one covariate by encoding each possible combination of them as a different value of one single covariate. For example, for two covariates being gender and nationality, we can encode X_i being one covariate with possible values female/American, male/Spanish and so forth. One could also consider real-valued covariates by cutting them into bins. Nevertheless, a future expansion should include the possibility to work with any type of metadata.

A community is a subset of vertices that share some properties. Formally, each node belongs to a community to an extent measured by a C -dimensional vector denoted *membership*. Since we are interested in directed networks, for each node i we assign two such vectors, u_i and v_i (for undirected networks we set $u = v$); these determine how i forms outgoing and incoming links respectively. Each layer α has an *affinity* matrix $W^{(\alpha)} = [w_{kl}^{(\alpha)}] \in \mathbb{R}^{C \times C}$ which describes the density of edges between each pair (k, l) of groups. Each community $k \in \{1, \dots, C\}$ is linked to a category $z \in \{1, \dots, Z\}$ by a parameter β_{kz} , that explains how much information of the z -th category is used to create the k -th community. To summarize, we consider two types of observed data: the adjacency tensor $A = \{A^{(\alpha)}\}_{1 \leq \alpha \leq L}$ and the design matrix $X = \{X_i\}_{i \in \{1, \dots, N\}}$; the first contains information about the networks topology structure, the latter about the node covariates. In addition, we have the model parameters that we compactly denote as $\Theta = \{U, V, W, \beta\}$.

The goal is to find the latent parameters Θ using the data A and X . In other words, given an observed multilayer network with adjacency tensor A and design matrix X , our goal is to simultaneously infer the node's membership vectors u_i and $v_i \forall i \in \{1, \dots, N\}$; the affinity matrices $W^{(\alpha)}, \forall \alpha \in \{1, \dots, L\}$, and the matrix $\beta = [\beta_{kz}] \in \mathbb{R}^{C \times Z}$, which captures correlations between communities and attributes. A visual overview of the proposed model is shown in Fig. 7. We consider a probabilistic generative model where MTCOV generates the network and the

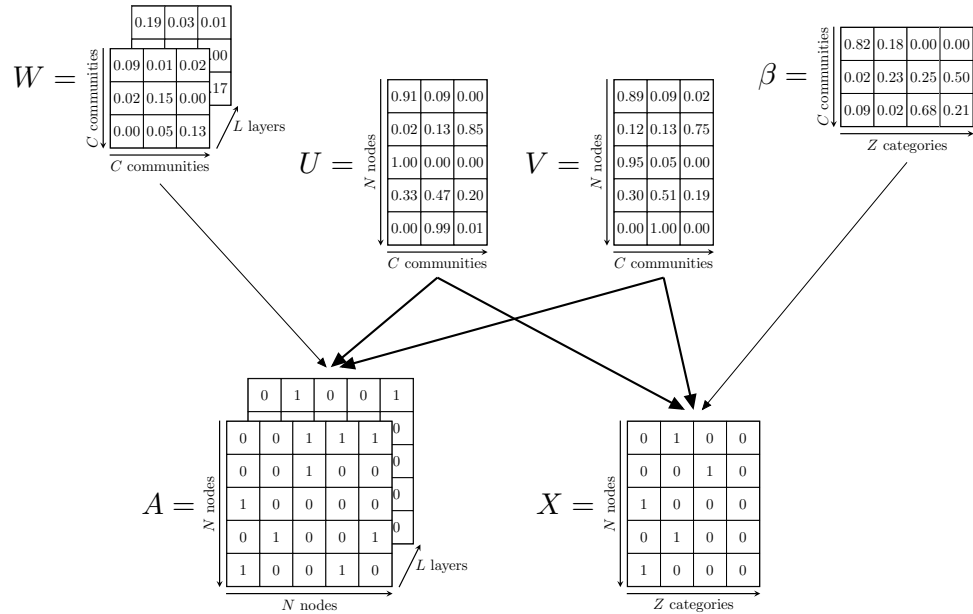


Figure 7. Graphical model representation of the algorithm MTCOV. A is the adjacency tensor, X is the design matrix and W, U, V, β are the latent parameters Θ . The membership matrices U and V couple the two datasets, and this is highlighted by the stronger thickness; whereas W and β are specific to the adjacency tensor and design matrix respectively. Here we present an example with binary adjacency matrix A , but the model is valid for more general weighted networks.

attributes probabilistically, assuming an underlying structure consisting of C overlapping communities. We adopt a maximum likelihood approach where, given the latent parameters Θ , we assume that the data A and X have independent likelihoods; in other words, we assume that A and X are *conditionally independent* given the latent parameters Θ . In addition, we assume that the memberships U and V couple the two datasets, as they are parameters shared between the two likelihoods; whereas the W and β are specific to the adjacency and design matrix respectively. We describe separately the procedures for modeling the topology of the network and the node attributes and then we show how to combine them in a unified log-likelihood framework.

Modeling the network topology. In modeling the likelihood of the network topology, we adopt the ideas behind MULTITENSOR: we assume that the expected number of edges of type α from i to j is given by the parameter:

$$M_{ij}^{(\alpha)} = \sum_{k,l=1}^C u_{ik} v_{jl} w_{kl}^{(\alpha)}. \tag{1}$$

We then assume that each entry $a_{ij}^{(\alpha)}$ of the adjacency tensor is extracted from a Poisson distribution with parameter $M_{ij}^{(\alpha)}$. This is a common choice for network data^{49–51} as it leads to tractable and efficient algorithmic implementations, compared for instance with other approaches that use Bernoulli random variables^{9,27}; it also allows the flexibility of treating both binary and integer-weighted networks. We further assume that, given the memberships and affinity matrices, the edges are distributed independently; this is again a conditional independence assumption.

We can then write the likelihood of the network topology as:

$$P_G(A|U, V, W) = \prod_{i,j=1}^N \prod_{\alpha=1}^L \frac{e^{-M_{ij}^{(\alpha)}} (M_{ij}^{(\alpha)})^{A_{ij}^{(\alpha)}}}{A_{ij}^{(\alpha)}!}, \tag{2}$$

which leads to the log-likelihood $\mathcal{L}_G(U, V, W)$ for the *structural dimension*:

$$\mathcal{L}_G(U, V, W) = \sum_{i,j,\alpha} \left[A_{ij}^{(\alpha)} \log \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} - \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} \right], \tag{3}$$

where we have neglected constants that do not depend on the parameters.

Modeling the node attributes. In modeling the likelihood of the attributes, we assume that this extra information is generated from the membership vectors; this captures the intuition that knowing a node’s com-

munity membership helps in predicting the value of the node’s attribute. This assumption has also been made in other models for single-layer attributed networks⁹ where one wants to enforce the tendency that nodes in the same community (for assortative structures) are likely to share common attributes. Different approaches^{37,38} assume instead independence between attributes and membership, which follows a different idea of observing an interaction between individuals if either they belong to the same community (for assortative structures) or they share an attribute or both.

Then, we model the probability of observing the z -th category for the attribute covariate of node i as the parameter:

$$\pi_{iz} = \frac{1}{2} \sum_{k=1}^C \beta_{kz} (u_{ik} + v_{ik}), \tag{4}$$

where β_{kz} is the probability of observing a particular category z together with a community k ; thus $\pi_i = (\pi_{i1}, \dots, \pi_{iZ})$ is a Z -dimensional vector such that $\pi_{iz} \in [0, 1]$ and $\sum_{z=1}^Z \pi_{iz} = 1, \forall i$. For convenience, we consider one-hot encoding for $x_i = (x_{i1}, \dots, x_{iZ})$, the realization of the random variable X_i ; $x_{iz} = 1$ if node i has attribute corresponding to category z , 0 otherwise and $\sum_{z=1}^Z x_{iz} = 1$; the original design matrix $X_{N \times 1}$ is thus translated into a binary matrix $X_{N \times Z}$.

We then assume that each entry X_i of the design matrix is extracted from a multinomial distribution of parameter π_i , which yields the likelihood of the covariates:

$$P_X(X_i = x_i | U, V, \beta) = P_X(X_{i1} = x_{i1}, \dots, X_{iZ} = x_{iZ} | U, V, \beta) = \pi_{i1}^{x_{i1}} \dots \pi_{iZ}^{x_{iZ}}. \tag{5}$$

In order to satisfy the sum constraint $\sum_{z=1}^Z \pi_{iz} = 1$, we impose the normalizations $\sum_{z=1}^Z \beta_{kz} = 1$, valid $\forall k$ and $\sum_{k=1}^C u_{ik} = \sum_{k=1}^C v_{ik} = 1$, valid $\forall i$. Such constraints are a particular case for which the general constraint for the multinomial parameter is satisfied. Although they are not the only choices, they allow us to give a probabilistic meaning to the components of β and the memberships U and V . As done for the network’s edges, we assume conditional independence for the attributes on the various nodes. This leads to the log-likelihood $\mathcal{L}_X(U, V, \beta)$ for the attribute dimension:

$$\mathcal{L}_X(U, V, \beta) = \sum_{i=1}^N \sum_{z=1}^Z x_{iz} \log(\pi_{iz}) = \sum_{i,z} x_{iz} \log\left(\frac{1}{2} \sum_k \beta_{kz} (u_{ik} + v_{ik})\right). \tag{6}$$

Note, we assume that the attributes have values that can be binned in a finite number Z of unordered categories and the attributes do not need to be one-dimensional. Indeed, we can encode each combination of more attributes as a different value of one-dimensional “super-attribute”. The model will not be affected, but the computational complexity might increase.

Inference with the EM algorithm. Having described how the model works and its main assumptions and intuitions, we now turn our attention to describe how to fit the parameters to the data, in other words, how to perform inference. We assume conditional independence between the network and attribute variables, thus we can decompose the total log-likelihood into a sum of two terms $\mathcal{L}(U, V, W, \beta) = \mathcal{L}_G(U, V, W) + \mathcal{L}_X(U, V, \beta)$. However, in practice, we can improve parameters’ inference performance by better balancing the contributions of the two terms as their magnitude can be on different scales, thus the risk of biasing the total likelihood maximization towards one of the two terms. For this, we introduce a scaling parameter $\gamma \in [0, 1]$ that explicitly controls the relative contribution of the two terms. The total log-likelihood is then:

$$\begin{aligned} \mathcal{L}(U, V, W, \beta) &= (1 - \gamma) \mathcal{L}_G(U, V, W) + \gamma \mathcal{L}_X(U, V, \beta) \\ &= (1 - \gamma) \sum_{i,j,\alpha} \left[A_{ij}^{(\alpha)} \log \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} - \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} \right] \\ &\quad + \gamma \sum_{i,z} x_{iz} \log\left(\frac{1}{2} \sum_k \beta_{kz} (u_{ik} + v_{ik})\right). \end{aligned} \tag{7}$$

Varying γ from 0 to 1 lets us interpolate between two extremes: analyzing the data purely in terms of the network topology or purely in terms of the attribute information. One can either fix this *a priori* based on the goal of the application, closer to 0 for link prediction or closer to 1 for attribute classification, or this can be treated as a hyperparameter that must be estimated, whose optimal value is obtained by fitting the data *via* tuning techniques (for instance cross-validation). This approach provides a natural quantitative measure for the dependence between the communities and the two sources of information. Notice that one can rescale *a priori* each likelihood term individually in order to control even more their magnitudes, and then add it to Eq. (7). This choice should be made based on the dataset at hand. Here we consider rescaling \mathcal{L}_G and \mathcal{L}_X only in studying the social support networks of Indian villages, as we have enough data for estimating the normalization coefficients; see Supplementary Section S3.1 for details.

We wish to find the $\Theta = (U, V, W, \beta)$ that maximizes Eq. (7). In general, this is computationally difficult, but we make it tractable by adopting a variational approach using an Expectation-Maximization (EM) algorithm⁵², similar to what done by De Bacco et al.¹³, but extended here to include attribute information. Namely, we introduce two probability distributions: h_{ikz} and $\rho_{ijkl}^{(\alpha)}$. For each i, z with $X_{iz} = 1$, h_{ikz} represents our estimate of the

probability that the i -th node has the z -th category, given that it belongs to the community k . On the other hand, for each i, j, α with $A_{ij}^{(\alpha)} = 1$, $\rho_{ijkl}^{(\alpha)}$ is the probability distribution over pairs of groups k, l .

Using Jensen's inequality $\log \bar{x} \geq \overline{\log x}$ for each log-likelihood term gives:

$$\begin{aligned} \mathcal{L}_X(U, V, \beta) &\geq \sum_{i,z} x_{iz} \sum_k h_{izk} \log \frac{\beta_{kz}(u_{ik} + v_{ik})}{2h_{izk}} = \sum_{i,z,k} x_{iz} [h_{izk} \log \beta_{kz}(u_{ik} + v_{ik}) - h_{izk} \log 2h_{izk}] \\ &= \mathcal{L}_X(U, V, \beta, h) \end{aligned} \tag{8}$$

$$\mathcal{L}_G(U, V, W) \geq \sum_{i,j,k,l,\alpha} \left[A_{ij}^{(\alpha)} \left(\rho_{ijkl}^{(\alpha)} \log u_{ik} v_{jl} w_{kl}^{(\alpha)} - \rho_{ijkl}^{(\alpha)} \log \rho_{ijkl}^{(\alpha)} \right) - u_{ik} v_{jl} w_{kl}^{(\alpha)} \right] = \mathcal{L}_G(U, V, W, \rho). \tag{9}$$

These lower bounds hold with equality when

$$h_{izk} = \frac{\beta_{kz}(u_{ik} + v_{ik})}{\sum_{k'} \beta_{k'z}(u_{ik'} + v_{ik'})}, \quad \rho_{ijkl}^{(\alpha)} = \frac{u_{ik} v_{jl} w_{kl}^{(\alpha)}}{\sum_{k',l'} u_{ik'} v_{jl'} w_{k'l'}^{(\alpha)}}, \tag{10}$$

thus maximizing $\mathcal{L}_X(U, V, \beta)$ is equivalent to maximizing $\mathcal{L}_X(U, V, \beta, h)$; similarly for $\mathcal{L}_G(U, V, W)$ and $\mathcal{L}_G(U, V, W, \rho)$ (this was also the same result derived by De Bacco et al.¹³). Overall, we aim at maximizing $\mathcal{L}(U, V, W, \beta, h, \rho) = (1 - \gamma)\mathcal{L}_G(U, V, W, \rho) + \gamma \mathcal{L}_X(U, V, \beta, h)$, in analogy with what was done before. These maximizations can be performed by alternatively updating a set of parameters while keeping the others fixed. The EM algorithm performs these steps by alternatively updating h, ρ (Expectation step) and Θ (Maximization step); this is done starting from a random configuration until $\mathcal{L}(\Theta, h, \rho)$ reaches a fixed point. Calculating Eq. (10) represents the E-step of the algorithm. The M-step is obtained by computing partial derivatives of $\mathcal{L}(\Theta, h, \rho)$ with respect to the various parameters in Θ and setting them equal to zero. We add Lagrange multipliers $\lambda = (\lambda^{(\beta)}, \lambda^{(u)}, \lambda^{(v)})$ to enforce constraints:

$$\mathcal{L}'(\Theta, h, \rho, \lambda) = \mathcal{L}(\Theta, h, \rho) - \sum_k \lambda_k^{(\beta)} \left(\sum_{z=1}^Z \beta_{kz} - 1 \right) - \sum_i \lambda_i^{(u)} \left(\sum_{k=1}^C u_{ik} - 1 \right) - \sum_i \lambda_i^{(v)} \left(\sum_{k=1}^C v_{ik} - 1 \right). \tag{11}$$

For instance, focusing on the update for β_{zk} , setting the derivative with respect to it in Eq. (11) to zero and enforcing the constraint $\sum_{z=1}^Z \beta_{kz} = 1$ gives $\lambda_k^{(\beta)} = \gamma \sum_{i,z} x_{iz} h_{izk}$; plugging this back finally gives:

$$\beta_{kz} = \frac{\sum_i x_{iz} h_{izk}}{\sum_{i,z} x_{iz} h_{izk}}, \tag{12}$$

which is valid for $\gamma \neq 0$. Doing the same for the other parameters yields (see Supplementary Section S1 for details):

$$u_{ik} = \frac{\gamma \sum_z x_{iz} h_{izk} + (1 - \gamma) \sum_{j,l,\alpha} A_{ij}^{(\alpha)} \rho_{ijkl}^{(\alpha)}}{\gamma + (1 - \gamma) \sum_{j,\alpha} A_{ij}^{(\alpha)}} \tag{13}$$

$$v_{ik} = \frac{\gamma \sum_z x_{iz} h_{izk} + (1 - \gamma) \sum_{j,l,\alpha} A_{ji}^{(\alpha)} \rho_{jilk}^{(\alpha)}}{\gamma + (1 - \gamma) \sum_{j,\alpha} A_{ji}^{(\alpha)}} \tag{14}$$

$$w_{kl}^{(\alpha)} = \frac{\sum_{i,j} A_{ij}^{(\alpha)} \rho_{ijkl}^{(\alpha)}}{\sum_i u_{ik} \sum_j v_{jl}}, \tag{15}$$

where Eq. (15) is valid for $\gamma \neq 1$. The EM algorithm thus consists in randomly initializing the parameters Θ and then repeatedly alternating between updating h and ρ using Eq. (10) and updating Θ using Eqs. (12)–(15) until $\mathcal{L}(\Theta, h, \rho)$ reaches a fixed point. A pseudo-code is given in Algorithm 1. In general, the fixed point is a local maximum but we have no guarantees that it is also the global one. In practice, we run the algorithm several times, starting from different random initializations and taking the run with the largest final $\mathcal{L}(\Theta, h, \rho)$. The computational complexity per iteration scales as $O(M^2 C^2 + NCZ)$, where M is the total number of edges summed across layers. In practice, C and Z have similar order of magnitude, usually much smaller than the system size M ; for sparse networks, as is often the case for real datasets, $M \propto N$, thus the algorithm is highly scalable with a total running time linear in the system size. An experimental analysis of the computational time is provided in the Supplementary Section S2.

Notice that, although we started from a network log-likelihood $\mathcal{L}_G(U, V, W)$ similar to the one proposed in the MULTITENSOR model¹³, the only update preserved from that is the one of w_{kl} in Eq. (15). The updates for u_{ik} and v_{ik} are instead quite different; the main reason is that here we incorporated the node attributes, which appear both explicitly and implicitly (through h) inside the updates. In addition, here we enforce normalizations like $\sum_k u_{ik} = 1$, not enforced in MULTITENSOR. This implies that our model restricted to $\gamma = 0$, i.e., no attribute information, does not correspond exactly to MULTITENSOR. This also implies that, upon convergence, we can

directly interpret the memberships as *soft* community assignments (or overlapping) without the need of post-processing their values; in words, u_{ik} represent the probability of node i to belong to the *outgoing* community k , similarly for v_{ik} and an *incoming* membership. This distinction is necessary when considering directed networks. If one is interested in recovering *hard* memberships, where a node is assigned to only one community, then one can choose the community corresponding to the maximum entry of u or v .

Algorithm 1 MTCOV- EM algorithm

Input: network $A = \{A_{ij}\}_{i,j=1}^N$, design matrix $X = \{x_{iz}\}_{i=1}^N$, number of communities C , hyperparameter γ

Output: membership vectors $U = [u_{ik}]$, $V = [v_{ik}]$; network-affinity matrix $W = [w_{kl}]$; attribute-affinity matrix $\beta = [\beta_{kz}]$.

Initialize U, V, W, β at random.

Repeat until convergence:

1. Calculate h and ρ (E-Step):

$$h_{izk} = \frac{\beta_{kz}(u_{ik} + v_{ik})}{\sum_{k'} \beta_{k'z}(u_{ik'} + v_{ik'})}, \quad \rho_{ijkl}^{(\alpha)} = \frac{u_{ik} v_{jl} w_{kl}^{(\alpha)}}{\sum_{k', l'} u_{ik'} v_{jl'} w_{k'l'}^{(\alpha)}}$$

2. Update parameters Θ (M-Step):

- i) for each node i and community k update memberships:

$$u_{ik} = \frac{\gamma \sum_z x_{iz} h_{izk} + (1 - \gamma) \sum_{j, l, \alpha} A_{ij}^{(\alpha)} \rho_{ijkl}^{(\alpha)}}{\gamma + (1 - \gamma) \sum_{j, \alpha} A_{ij}^{(\alpha)}}$$

$$v_{ik} = \frac{\gamma \sum_z x_{iz} h_{izk} + (1 - \gamma) \sum_{j, l, \alpha} A_{ji}^{(\alpha)} \rho_{jilk}^{(\alpha)}}{\gamma + (1 - \gamma) \sum_{j, \alpha} A_{ji}^{(\alpha)}}$$

- ii) if $\gamma \neq 1$, for each pair of communities (k, l) update network-affinity matrix:

$$w_{kl}^{(\alpha)} = \frac{\sum_{i, j} A_{ij}^{(\alpha)} \rho_{ijkl}^{(\alpha)}}{\sum_i u_{ik} \sum_j v_{jl}}$$

- iii) if $\gamma \neq 0$, for each pair of community-attribute (k, z) update attribute-affinity matrix:

$$\beta_{kz} = \frac{\sum_i x_{iz} h_{izk}}{\sum_{i, z} x_{iz} h_{izk}}$$

Evaluation metrics. We adopt two different criteria for performance evaluation, based on having or not having access to ground-truth values for the community assignments. The first case applies to synthetic-generated data, the second to both synthetic and real-world data. We explain performance metrics in detail below.

Ground-truth available. In the presence of a known partition, we measure the agreement between the set of ground-truth communities \mathcal{C}^* and the set of detected communities \mathcal{C} using metrics for recovering both hard and soft assignments. For hard partitions, the idea is to match every detected community with its most similar ground-truth community and measure similarity $\delta(\mathcal{C}_i^*, \mathcal{C}_j)$ (and vice versa for every ground-truth community matched against a detected community) as done by Yang et al.⁹. The final performance is the average of these two comparisons:

$$\frac{1}{2|\mathcal{C}^*|} \sum_{\mathcal{C}_i^* \in \mathcal{C}^*} \max_{\mathcal{C}_j \in \mathcal{C}} \delta(\mathcal{C}_i^*, \mathcal{C}_j) + \frac{1}{2|\mathcal{C}|} \sum_{\mathcal{C}_j \in \mathcal{C}} \max_{\mathcal{C}_i^* \in \mathcal{C}^*} \delta(\mathcal{C}_i^*, \mathcal{C}_j), \tag{16}$$

where here we consider as similarity metric $\delta(\cdot)$ the F1-score and the Jaccard similarity.

In both cases, the final score is a value between 0 and 1, where 1 indicates the perfect matching between detected and ground-truth communities. For soft partitions, we consider two standard metrics for measuring distance between vectors as done by De Bacco et al.¹³, such as *cosine similarity* (CS) and L_1 error, averaged over the nodes:

$$CS(U, U^0) = \frac{1}{N} \sum_{i=1}^N \frac{u_i \cdot u_i^0}{\|u_i\|_2 \|u_i^0\|_2} = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C \frac{u_{ik} u_{ik}^0}{\|u_i\|_2 \|u_i^0\|_2} \tag{17}$$

$$L_1(U, U^0) = \frac{1}{2N} \sum_{i=1}^N \|u_i - u_i^0\|_1 = \frac{1}{2N} \sum_{i=1}^N \sum_{k=1}^C |u_{ik} - u_{ik}^0|, \quad (18)$$

where u_i is the C -dimensional vector containing the i -th row of U , representing the detected membership and similarly for u_i^0 for the ground-truth U^0 . The factor $1/2$ ensures that the L_1 distance ranges from 0 for identical distributions to 1 for distributions with disjoint support. Similarly to the what done for hard partitions, we match the ground-truth and detected communities by choosing the permutation of C groups that gives the highest cosine similarity or smallest L_1 distance.

Ground-truth not available. In the absence of ground-truth, these metrics cannot be computed, and one must resort to other approaches for model evaluation. Here we consider performance in prediction tasks when hiding part of the input datasets while fitting the parameters, and in particular on the extent to which partial knowledge of network edges helps predict node attributes and vice versa. Thus we consider a measure for link-prediction and one for correct retrieval of the attributes. For link-prediction, we used the AUC statistic, equivalent to the area under the receiver-operating characteristic (ROC) curve⁵³. It represents the probability that a randomly chosen missing connection (a true positive) is given a higher score than a randomly chosen pair of unconnected vertices (a true negative). Thus, an AUC statistic equal to 0.5 indicates random chance, while the closer it is to 1, the more our model's predictions are better than chance. We measure the probability of observing an edge as the predicted expected Poisson parameters of Eq. (1). For the attribute, instead, we use the accuracy as a quality measure. For each node, we compute the predicted expected multinomial parameter π_i using Eq. (4). We then assign to each node the category with the highest probability, computing the accuracy as the ratio between the correctly classified examples over the total number of nodes. As baselines, we compare with the accuracy obtained with a random uniform probability and the highest relative frequency observed in the training set.

Cross-validation tests and hyperparameter settings. We perform prediction tasks using cross-validation with 80–20 splits: we use 80% of the data for training the parameters and then measure AUC and accuracy on the remaining 20% test set. Specifically, for the network topology, we hold out 20% of the triples (i, j, α) ; for the attributes, we hold out 20% of the entries of the categorical vector.

Our model has two hyperparameters, the scaling parameter γ and the number of communities C . We estimate them by using 5-fold cross-validation along with grid search to range across their possible values. We then select the combination $(\hat{C}, \hat{\gamma})$ that returns the best average performance over the cross-validation runs. Standard cross-validation considers performance in terms of a particular metric. However, here we have two possible ones which are qualitatively different, i.e., AUC and accuracy. Depending on the task at hand, one can define performance as a combination of the two, bearing in mind that the values of $(\hat{C}, \hat{\gamma})$ at the maximum of either of them might not coincide. Here we select $(\hat{C}, \hat{\gamma})$ as the values are jointly closer to both the maximum values. In the experiments where one of the two hyperparameters is fixed *a priori*, we run the same procedure but vary with grid search only the unknown hyperparameter.

Data availability

The code used for the analysis and to generate the synthetic data is publicly available and can be found at <https://github.com/mcontisc/MTCOV>.

Code availability

An open-source algorithmic implementation available at <https://github.com/mcontisc/MTCOV>.

Received: 28 April 2020; Accepted: 4 September 2020

Published online: 25 September 2020

References

1. Waskiewicz, T. Friend of a friend influence in terrorist social networks. In *Proceedings on the international conference on artificial intelligence (ICAI)*, 1 (The Steering Committee of The World Congress in Computer Science, Computer..., 2012).
2. Pinheiro, C. A. R. Community detection to identify fraud events in telecommunications networks. In *SAS SUGI proceedings: customer intelligence* (2012).
3. Pan, W.-E., Jiang, B. & Li, B. Refactoring software packages via community detection in complex software networks. *Int. J. Autom. Comput.* **10**, 157–166 (2013).
4. Bechtel, J. J. *et al.* Lung cancer detection in patients with airflow obstruction identified in a primary care outpatient practice. *Chest* **127**, 1140–1145 (2005).
5. Chen, J., Zhang, H., Guan, Z.-H. & Li, T. Epidemic spreading on networks with overlapping community structure. *Physica A Stat. Mech. Appl.* **391**, 1848–1854 (2012).
6. Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53**, 526–543 (2011).
7. Newman, M. E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**, 8577–8582 (2006).
8. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
9. Yang, J., McAuley, J. & Leskovec, J. Community detection in networks with node attributes. In *2013 IEEE 13th international conference on data mining*, 1151–1156 (IEEE, 2013).
10. Falih, I., Grozavu, N., Kanawati, R. & Bennani, Y. Community detection in attributed network. *Companion Proc. Web Conf.* **2018**, 1299–1306 (2018).

11. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
12. De Domenico, M. *et al.* Mathematical formulation of multilayer networks. *Phys. Rev. X* **3**, 041022 (2013).
13. De Bacco, C., Power, E. A., Larremore, D. B. & Moore, C. Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
14. Schein, A., Paisley, J., Blei, D. M. & Wallach, H. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, 1045–1054 (2015).
15. Schein, A., Zhou, M., Blei, D. M. & Wallach, H. Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *Proceedings of the 33rd international conference on machine learning*, vol. 48 (2016).
16. Valles-Catala, T., Massucci, F. A., Guimera, R. & Sales-Pardo, M. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X* **6**, 011036 (2016).
17. Stanley, N., Shai, S., Taylor, D. & Mucha, P. Clustering network layers with the strata multilayer stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **3**, 95–105 (2016).
18. Peixoto, T. P. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E* **92**, 042807 (2015).
19. Paul, S. *et al.* Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Stat.* **10**, 3807–3870 (2016).
20. Gheche, M. E., Chierchia, G. & Frossard, P. Orthonet: multilayer network data clustering. *IEEE Trans. Signal Inf. Process. Netw.* **6**, 13–23 (2020).
21. Papadopoulos, A., Rafailidis, D., Pallis, G. & Dikaiakos, M. D. Clustering attributed multi-graphs with information ranking. In *Proceedings, Part I, of the 26th international conference on database and expert systems applications—volume 9261, DEXA 2015*, 432–446 (Springer, 2015).
22. Papadopoulos, A., Pallis, G. & Dikaiakos, M. D. Weighted clustering of attributed multi-graphs. *Computing* **99**, 813–840 (2017).
23. Chang, S. *et al.* Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '15*, 119–128 (2015).
24. Sachan, M., Contractor, D., Faruque, T. A. & Subramaniam, L. V. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on world wide web, WWW '12*, 331–340 (2012).
25. Sweet, T. M. & Zheng, Q. Estimating the effects of network covariates on subgroup insularity with a hierarchical mixed membership stochastic blockmodel. *Soc. Netw.* **52**, 100–114 (2018).
26. Signorelli, M. & Wit, E. C. Model-based clustering for populations of networks. *Stat. Model.* **20**, 9–29 (2019).
27. Newman, M. E. & Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).
28. Bothorel, C., Cruz, J. D., Magnani, M. & Micenkova, B. Clustering attributed graphs: models, measures and methods. *Netw. Sci.* **3**, 408–444 (2015).
29. Zhang, Y. *et al.* Community detection in networks with node features. *Electron. J. Stat.* **10**, 3153–3178 (2016).
30. Hric, D., Peixoto, T. P. & Fortunato, S. Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X* **6**, 031038 (2016).
31. Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M. & Mucha, P. J. Stochastic block models with multiple continuous attributes. *Appl. Netw. Sci.* **4**, 1–22 (2019).
32. Emmons, S. & Mucha, P. J. Map equation with metadata: varying the role of attributes in community detection. *Phys. Rev. E* **100**, 022301 (2019).
33. Xu, Z., Ke, Y., Wang, Y., Cheng, H. & Cheng, J. A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, 505–516 (2012).
34. Bu, Z., Li, H.-J., Cao, J., Wang, Z. & Gao, G. Dynamic cluster formation game for attributed graph clustering. *IEEE Trans. Cybern.* **49**, 328–341 (2017).
35. Tallberg, C. A Bayesian approach to modeling stochastic blockstructures with covariates. *J. Math. Sociol.* **29**, 1–23 (2004).
36. White, A. & Murphy, T. B. Mixed-membership of experts stochastic blockmodel. *Netw. Sci.* **4**, 48–80 (2016).
37. Airoldi, E. M., Choi, D. S. & Wolfe, P. J. Confidence sets for network structure. *Stat. Anal. Data Min. ASA Data Sci. J.* **4**, 461–469 (2011).
38. Sweet, T. M. Incorporating covariates into stochastic blockmodels. *J. Educ. Behav. Stat.* **40**, 635–664 (2015).
39. Taylor, D., Shai, S., Stanley, N. & Mucha, P. J. Enhanced detectability of community structure in multilayer networks through layer aggregation. *Phys. Rev. Lett.* **116**, 228301 (2016).
40. Taylor, D., Caceres, R. S. & Mucha, P. J. Super-resolution community detection for layer-aggregated multilayer networks. *Phys. Rev. X* **7**, 031056 (2017).
41. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: first steps. *Soc. Netw.* **5**, 109–137 (1983).
42. Power, E. A. *Building Bigness: Religious Practice and Social Support in Rural South India*. Doctoral Dissertation, Stanford University, Stanford, CA (2015).
43. Power, E. A. Social support networks and religiosity in rural South India. *Nat. Hum. Behav.* **1**, 0057 (2017).
44. Power, E. A. & Ready, E. Cooperation beyond consanguinity: post-marital residence, delineations of kin and social support among South Indian Tamils. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20180070 (2019).
45. McAuley, J. & Leskovec, J. Learning to discover social circles in ego networks. In *Proceedings of the 25th international conference on neural information processing systems—volume 1, NIPS'12*, 539–547 (2012).
46. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).
47. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the 3rd international workshop on link discovery, LinkKDD '05*, 36–43 (2005).
48. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM Rev.* **51**, 455–500 (2009).
49. Ball, B., Karrer, B. & Newman, M. E. J. Efficient and principled method for detecting communities in networks. *Phys. Rev. E* **84**, 036103 (2011).
50. Gopalan, P. K. & Blei, D. M. Efficient discovery of overlapping communities in massive networks. *Proc. Natl. Acad. Sci. USA* **110**, 14534–14539 (2013).
51. Gopalan, P., Hofman, J. M. & Blei, D. M. Scalable recommendation with hierarchical Poisson factorization. In *Proceedings of the 31st conference on uncertainty in artificial intelligence*, 122–129 (2015).
52. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**, 1–22 (1977).
53. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

Acknowledgements

This work was partially supported by the Cyber Valley Research Fund. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani. The authors are grateful for the goodwill of the residents of Tenpaṭṭi and Alakapuram, the support of faculty and students from

the Folklore Department at Madurai Kamaraj University, and the assistance of the Chella Meenakshi Centre for Educational Research and Services. Funding for fieldwork was provided by the US National Science Foundation Doctoral Dissertation Improvement Grant (No. BCS-1121326), a Fulbright-Nehru Student Researcher Award, the Stanford Center for South Asia, and a National Science Foundation Interdisciplinary Behavioral & Social Science Research Grant (No. IBSS-1743019). We thank Cristopher Moore and Daniel Larremore for useful discussions and the Santa Fe Institute for providing the environment fostering these interactions.

Author contributions

M.C., E.P. and C.D.B. conceived the research and designed the analyses; M.C. conducted the experiment; all authors reviewed the manuscript.

Funding

Open Access funding provided by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-72626-y>.

Correspondence and requests for materials should be addressed to C.D.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships

Kyle Higham^{*,1,a}, Martina Contisciani^{2,b}, Caterina De Bacco^{3,b}

^a Institute of Innovation Research, Hitotsubashi University, Tokyo, Japan

^b Max Planck Institute for Intelligent Systems, Cyber Valley, Tübingen, Germany

ABSTRACT

The use of patent citation networks as research tools is becoming increasingly commonplace in the field of innovation studies. However, these networks rarely consider the contexts in which these citations are generated and are generally restricted to a single jurisdiction. Here, we propose and explore the use of a multilayer network framework that can naturally incorporate citation metadata and stretch across jurisdictions, allowing for a complete view of the global technological landscape that is accessible through patent data. Taking a conservative approach that links citation network layers through triadic patent families, we first observe that these layers contain complementary, rather than redundant, information about technological relationships. To probe the nature of this complementarity, we extract network communities from both the multilayer network and analogous single-layer networks, then directly compare their technological composition with established technological similarity networks. We find that while technologies are more splintered across communities in the multilayer case, the extracted communities match much more closely the established networks. We conclude that by capturing citation context, a multilayer representation of patent citation networks is, conceptually and empirically, better able to capture the significant nuance that exists in real technological relationships when compared to traditional, single-layer approaches. We suggest future avenues of research that take advantage of novel computational tools designed for use with multilayer networks.

1. Introduction

Patent citations have found successful application in a wide swathe of contexts, from understanding knowledge spillovers (Berkes and Gaetani, 2021; Jaffe and de Rassenfosse, 2017; Jaffe et al., 1993; Sorenson et al., 2006) to the characterization of technological change (Choi and Park, 2009; Fleming, 2001; Huenteler et al., 2016). The vast majority of this research is conducted using only patent data from a single jurisdiction and often ignores important citation context. However, as innovation and patent filings become increasingly global endeavours (Danguy, 2017; Fink et al., 2016), there are many situations where it is important to think of ‘the patent system’ as a set of quasi-coordinated processes operating across jurisdictional boundaries (Petit et al., 2021).

This coordination is desirable because the same invention can be patented in multiple jurisdictions; there are clear efficiency gains to be made if information discovered or produced during the patent prosecution process can be shared between jurisdictions (Chun, 2011). Patent families arise because these related applications, which simultaneously

progress through multiple patent offices, are legally linked through their first filing. The Paris Convention⁴ allows applicants to apply in multiple jurisdictions and claim the filing date of the first application, known as the priority date, as the effective filing date for subsequent applications (provided these occur within 12 months of the priority date). Information sharing between offices creates, by design, some redundancy in the information generated for family members across offices, but not enough that a complete picture can be pieced together from a single jurisdiction’s data. The existence of patent families provides the opportunity to form the most complete set of information about a particular invention that can be obtained from patent data and allows us to link metadata across jurisdictional boundaries (Nakamura et al., 2015). In this work, we demonstrate the utility of these linkages in the context of patent citation networks.

The family-level view would suggest that only using data from a single jurisdiction leaves a lot of potentially relevant information unexamined (Bakker et al., 2016). In the more and more common scenario where multiple family members exist across multiple jurisdictions,

* Corresponding author.

E-mail addresses: higham@iir.hit-u.ac.jp (K. Higham), martina.contisciani@tuebingen.mpg.de (M. Contisciani), caterina.debacco@tuebingen.mpg.de (C. De Bacco).

¹ Postdoctoral Fellow

² PhD student

³ Independent Research Group Leader

⁴ Paris Convention for the Protection of Industrial Property (1883).

citations will often only be made to one family member.⁵ As such, the citation network that is obtained from any single jurisdiction necessarily represents a subset of the complete network for the set of inventions under examination. While information sharing between offices will increase the amount of overlap between these networks, it does not make family-level analyses redundant, for two reasons. First, the amount of information sharing, and the modes for doing so, between patent offices has changed significantly in recent years.⁶ In particular, advances in information technology allow patent offices to coordinate much more effectively than they did 20 years ago. Yet, some patents filed 20 years ago are only expiring now, so these patents can still be important sources of information when studying contemporary innovation. Second, many patents are *not* filed in multiple offices, and some applicants only select a few strategically important jurisdictions where they would like to protect their intellectual property. That is, the nodes and links in the citation networks of each jurisdiction are unique, and so each network contains a huge amount of potentially pertinent information that is unique to that jurisdiction. In the empirical sections of this work, we take a very conservative approach and only consider nodes that are shared across jurisdictions, as described in detail in Section 2.3.

Even for shared nodes (defined here as patents granted in multiple jurisdictions), however, the sets and types of citations made by each patent office can differ greatly, as shown in Fig. 1. The primary reason for this disagreement is that different jurisdictions abide by different legal guidelines that describe when and how citations should be made. These sets of guidelines are not without strong similarities, however, and a careful reading offers pathways towards sensible aggregation or comparison of these sets of citations (Higham and Yoshioka-Kobayashi, 2022). This has become particularly feasible in recent times as more and more offices now provide metadata about citation context, such as whether the cited patent was so similar to the application as to render the latter unpatentable, or whether the cited patent was added to simply define the state of the art. A secondary reason for disagreement between jurisdictions is, in fact, a commonality: examiners in all jurisdictions are humans with limited time to examine any particular patent (see, e.g., Frakes and Wasserman, 2017). Often, it is simply not possible to find every relevant piece of prior art, particularly when language barriers are taken into consideration. Indeed, in combination with simple differences of opinion, this limitation means it is unlikely that two examiners in the same patent office would find exactly the same set of prior art (Wada, 2016). Therefore, using family-level information gives us the search result of more examiner-hours as well as the multiple opinions of what should be considered relevant prior art.

As citations made by different offices are made according to different sets of guidelines, treating these citations as equally informational may lead to misleading results. Indeed, some suggest that citations of the *same* type have become less informative over time (Kuhn et al., 2020). It is therefore important to aggregate family-level information sensibly. We propose that a multilayer network framework provides a natural representation of the patent citation network that readily incorporates differences in citation type. After all, multiple networks anchored by common nodes is the very definition of a multilayer network (De Domenico et al., 2013; Kivela et al., 2014; Porter, 2018).

Within the multilayer framework, described in more detail for our chosen context in Section 2.2, each layer of the network represents a single link type, each node represents a patent family (which may exist in multiple layers), and each link represents a citation between families (of the type defined by the layer). As such, the global patent citation network is an inherently multilayer system; no abstraction is required. Further, this framework is particularly flexible. For example, layers can

represent jurisdictions, and links within each jurisdictional layer can represent the citations found on the front page(s) of the family member (s) granted by that jurisdiction. From this point, it is possible to layer as many jurisdictions as desired onto the network, provided there are family linkages existing between the layers. It is also possible to split these layers further, according to citation metadata that inform us of the reason for, or source of, a particular citation. This flexibility is particularly valuable when certain types of citations are irrelevant, or may even be considered pure noise, with respect to a particular research question. For example, one studying knowledge flow within a multilayer framework may not wish to consider citations discovered by the examiner, and may even want to add an additional layer for citations found in the patent specification (Verluse et al., 2020).

However, it is not clear, a priori, whether a multilayer framework adds any information over and above that which can be found in ‘flattened’ family citation networks wherein citation context is disregarded and only link existence is examined (Nakamura et al., 2015). Thus, in order for the multilayer framework to be feasible as a research tool, it is important to first demonstrate a significant gain in information content relative to the flattened, global family citation network, or even the more commonly-used jurisdictionally restricted citation networks. To this end, we explore the information content of the triadic patent family network, wherein all layers contain the same set of nodes. This set consists of families containing at least one member granted in each of the triadic patent offices: the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), and the Japan Patent Office (JPO). The triadic offices have historically granted the majority of patents globally and contain rich and accessible citation information. Specifics about the data used in this work can be found in Section 2.3.

In this work, we first construct a multilayer family-family triadic citation network, wherein layers can be separated by jurisdiction and citation context (such as whether the citation was added by an applicant or examiner). In practice, the appropriate set of contexts can be selected based on the use-case; in this work, the additional context we consider is whether a citation was likely to have been found by the examiner or by the applicant (which is not always explicit), as we expect the differing motivations for citation between these groups affect the nature of the technological relationships reflected by these citations. We then conduct an interdependence analysis to check for redundancy of information content between the layers, finding that significant complementary information exists between jurisdictions. A community detection procedure is then conducted on the multilayer network and two comparison networks: the flattened multilayer network containing the same set of links but without information about jurisdiction or citation context, and the US-only subset of the citation network, also flattened. The former comparison tests the role of citation context, while the latter is included as the most commonly used patent citation network in prior research. We observe, graphically, nuanced differences in inferred community structure between the multilayer network and the comparison networks.

To add colour to these differences, we examine the relationships between inferred community partitions and the technology classes of the families that comprise them. For the multilayer network communities and those of the two flattened comparison networks, we project the bipartite community-class network onto the class nodes and directly compare these projections with established class-class networks (co-classification and inter-class citation linkage) with known-node-correspondence methods. We are also able to directly measure the diversity of communities, and the spread of classes between communities to inform our interpretation of the direct network comparisons.

When compared to the other two networks, we find that the multilayer case produces communities that more closely reflect the known technological relationships implied by the established class-class networks, at both micro- and meso-scales. Further, while technological classes are more splintered across communities in the multilayer case, the internal diversity of communities is lower than the comparison networks once we account for the known technological similarity of

⁵ Search reports will often list equivalents of the prior art that is cited, however, this additional information is not explicitly included in the associated data sets.

⁶ See, e.g., <https://www.wipo.int/case/en/>.

classes. These results suggest that, even within our conservative empirical framework, citation context is an important source of information about the nature and importance of the particular technological relationships codified by citation linkages, and that examination of multilayer citation networks using novel computational techniques is an exciting and relevant avenue for future research.

The rest of the paper is structured as follows. Section 2 introduces both patent families and multilayer networks and discusses how the former naturally forms the latter in the context of citation networks. Section 2.3 describes the data we use in this work, how this forms the multilayer networks and why specific subsets of families and citations are selected for analysis. Section 3 describes the empirical procedures that we use to test and compare the information content of the multilayer citation network relative to single-layer networks and describes the results obtained. Lastly, Section 4 concludes and discusses the limitations and extensions of this research.

2. Multilayer patent networks

2.1. Patent Families

The rights bestowed by patents are only enforceable in the jurisdiction in which the patent was granted. To obtain these rights in more than one jurisdiction, an applicant first files in a single jurisdiction (often their local patent office), starting the clock on the period during which they can file for the same invention in other jurisdictions. For the next 12 months, all subsequent filings can ‘claim priority’ from this initial application and inherit the latter’s filing date as its own for the purposes of examination (provided the same content is covered in the application).

There are two primary modes through which an invention can claim priority from an earlier application: the Paris Convention and the Patent Cooperation Treaty (PCT). The former lays down the guidelines for the treatment of foreign patent applications among the contracting parties, including the time limit on priority claims as described above. The latter, for our purposes here, is effectively an attempt to streamline and harmonise the process of patenting in multiple jurisdictions.⁷ This process does not result in a patent, but rather a preliminary prior art search report, and allows the applicant to nominate the jurisdictions to which they would like to apply for a patent without having to apply at each office separately. Priority can be claimed from a PCT filing, and PCT filings can themselves claim priority from an earlier filing at a local office.

After a patent application has reached a local office, the applicant may want to fine-tune their claims or even be asked to split the described invention into two separate patent applications.⁸ The inventor is not able to disclose new information during this process, and thus the claims made by the ‘new version’ of the application must be contained within the scope of the initial disclosure. These subsequent filings may claim the priority date of the initial filings and are referred to as ‘continuing applications’.

Patent families, in general, link patents and applications through their priority filing. The resulting ‘family trees’ can be complex and, as such, several types of families exist (Martinez, 2010; Martínez, 2011). ‘Simple’ patent families (as defined by the EPO for their DOCDB database) each consist of a set of patents and applications that are all linked to the same priority filing. This type of family is the one on which we focus in this work, and we will henceforth drop ‘simple’. As such, families can be made up of sets of documents from several jurisdictions, each of which may contain multiple documents. Other families may only consist of a single application in a single jurisdiction.

Families are the unit of analysis for the current work for two reasons. First, they generally align with what one usually thinks of as a single ‘invention’ (Martínez, 2011) and it is the relationships between inventions that we usually aim to capture with citation data. Second, they link inventions across jurisdictions, and therefore allow the alignment of jurisdiction-specific citation networks and, therefore, the introduction of the multilayer network as a potentially useful analytical, conceptual, and mathematical tool to study technological relationships.

From the perspective of data availability, detail, and volume, the obvious choice of data set for testing the utility of the multilayer framework are those patents granted by the three (historically) largest patent offices, also known as the triadic offices: the USPTO, EPO, and JPO. Further, we wish to take a particularly conservative empirical approach to these initial explorations of the multilayer citation network. To do this, we only consider patent families that have granted members in all layers of interest (‘triadic families’) and only consider citations among these families.⁹

Theoretically, each office examining these triadic applications has access to the same information regarding prior art, and they share much of what they find with the other two offices, directly or indirectly (Petit et al., 2021; Wada, 2020). For this reason, granted members have all had the same opportunities to link to other (older) families in each layer, maximising potential redundancies between layers in the citation network. The exclusion of families that are not triadic, therefore, is why we think of this analysis as likely to produce very conservative results when compared to those that may be obtained for a network without such exclusions.

We also note triadic patent families are often used as a binary indication of a ‘high-quality’ invention (de Rassenfosse and van Pottelsberghe, 2009; Tahmooresnejad and Beaudry, 2019); after all, the applicants thought it was worth the time and money to patent their invention in three of the largest markets in the world. By this logic, our multilayer network consists exclusively of ‘high-quality’ patent families¹⁰ and excludes much controversial subject matter that are not universally patentable (Biddinger, 2000).

A simplified diagram of a multilayer network of citations between triadic families is shown in Fig. 1, with full details of the families included in this example described in Appendix C. Note that the multilayer network that we analyse in this paper treats sub-layers, such as whether a citation has been used in a rejection decision (shown in red in Fig. 1), as distinct layers. This results in seven layers in total, as the EPO also provides information about whether a citation originated from the international search report (conducted outside the EPO) or the local search report.

2.2. Multilayer networks

Multilayer networks have received particular attention in the past decade (Boccaletti et al., 2014; Cimini et al., 2019; De Domenico et al., 2013; Kivelä et al., 2014), and the development of mathematical and computational tools for their analysis, as well as their timely application, remains a very active field of research across many domains (Gallotti et al., 2016; Harvey et al., 2021; van der Marel et al., 2021; Vaiana and Muldoon, 2020; Yuvaraj et al., 2021). In this work, we not only suggest that patent citation networks are naturally multilayered, but aim to introduce the multilayer framework to the innovation studies community to promote the timely application of novel computational tools that are currently being developed.

To date, the vast majority of the studies that explicitly place patent citation data into a network setting use a single-layer framework

⁹ The applicants to these offices, however, may be based outside these jurisdictions.

¹⁰ Note that this is a very narrow view of patent quality. For a comprehensive discussion, refer to Higham et al. (2021).

⁷ <https://www.wipo.int/pct/en/>.

⁸ See, e.g., Paris Convention for the Protection of Industrial Property (1883), Article 4G.

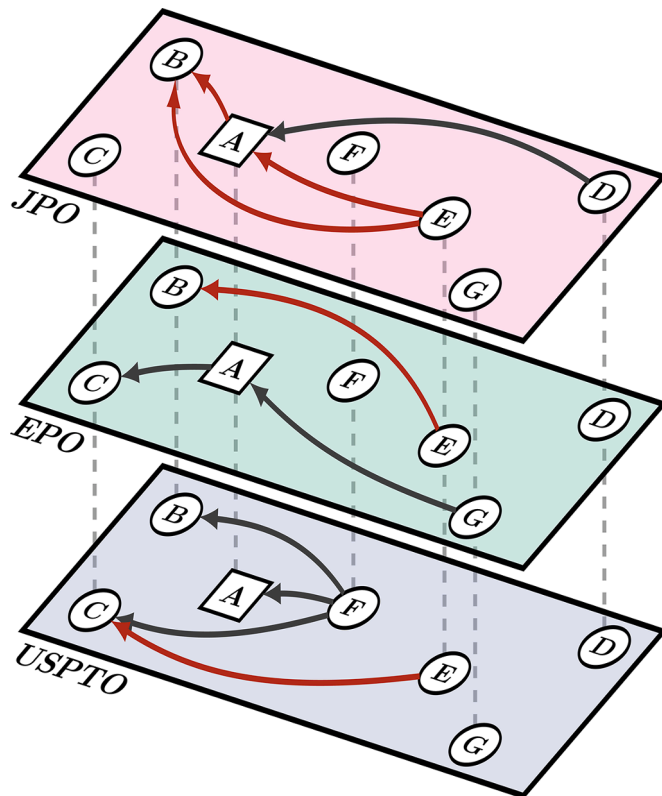


Fig. 1. Exemplar subset of the multilayer patent citation network. A multilayer representation of a typical subset of the inter-family patent citation network we consider in this work. Nodes and links comprise the multilayer ego network of patent family A, the USPTO equivalent of which is “Power source apparatus” (US6819081B2), initially filed in January 2002 by Sanyo Electric Co., Ltd. at the JPO. Each layer represents the inter-family citations made by a different patent office, and red links are those used to justify a (non-final) rejection of the application that was examined in that layer. All data represented here is subject to the restrictions described in Section 2.3 and is, therefore, an extremely simplified version of the complete ego network. Details of the families represented can be found in Appendix C.

(Clough et al., 2015; Funk and Owen-Smith, 2017; Higham et al., 2019; Mariani et al., 2019; Nakamura et al., 2015; Valverde et al., 2007; Von Wartburg et al., 2005; Wu et al., 2019). That is, there is only one type of link (i.e., a citation) between nodes in the network. This approach often makes practical sense, such as when one lacks citation metadata that may be used to distinguish or ‘colour’ the links, or if only one link type is of interest. However, a multilayer network framework is able to naturally incorporate citation metadata, if it exists, into the network structure.¹¹ As an analogy, let us consider the public transport network of a large city containing several different forms of transport, each with its own network of routes and stations. There are usually many points of overlap between these network layers to allow passengers to transfer between modes of public transport, such as a bus stop at a train station. These transfer points link the different network layers together. From both mathematical and computational perspectives, this kind of network is fundamentally different from single-layered networks, particularly when the different layers are defined by links with very different properties (Aleta et al., 2017; Ibrahim et al., 2021). In the public transport context, these properties can be straightforward, such as

¹¹ In a related work, also using triadic patents, Morrison et al. (2014) use a multiplex PageRank to assess the centrality of technology classes where layers are defined by inventor location. However, citation source and context are not considered.

speed, price, comfort, or environmental harm, or more computationally complex, such as sensitivity to link removal and amenability to rerouting (De Domenico et al., 2014).

In the domain of patent citation networks, each jurisdiction has a set of applications and patents that each contain a set of citations made to other documents. Each of those citations comes with context (Higham and Yoshioka-Kobayashi, 2022). This context can be whether the prior art was discovered by the examiner, the justification for its addition to the document, the relationship between the citing and cited firms, or any other citation metadata that may be obtained or constructed. For many research questions that rely on information derived from the citation network, this information is important to retain, just as it is important to know whether two nodes in a transport network are connected by a bus, an airplane, or a ferry.

At the same time, every patent is part of a family (even if there is only one member). When families contain members filed in multiple jurisdictions, the citation networks associated with each jurisdiction can be linked, just as a bus may stop at a train station, or a train may stop at an airport. Of course, patent applicants are under no obligation to file for a patent on the same invention in multiple jurisdictions. That is, a node (patent family) may not exist in all layers of the network. Not every bus stop is a train station, nor vice versa. The full patent citation network is a true ‘multilayer’ network in this sense. In this work, however, we focus on the subset of nodes that exist across all three layers of interest (the triadic offices). The justifications for this choice are discussed in Section 2.3. The network we define in this work, therefore, is a special case of a multilayer network wherein the layers are node-aligned (Kivela et al., 2014). Extensions of this work to a more general multilayer framework are discussed in Section 4.

Multilayer networks share many characteristics of interest that are found in single-layer networks; indeed, much of the early research on multilayer networks involved adapting concepts from single-layer networks to this new framework (Battiston et al., 2014; Berlingerio et al., 2011; Bródka et al., 2012; De Domenico et al., 2013). For our purposes, in order to demonstrate the utility of the multilayer framework, it is necessary to compare the network properties derived in this setting to those obtained from the equivalent, flattened single-layer network, wherein citation metadata is ignored (partially or wholly).

The domain within which we choose to explore differences between the multilayer and single-layer frameworks, in the patent citation context, is community detection. The natural grouping of nodes is one of the characteristic features of real-world networks and plays a significant role in describing the structure of the network at scales between node-level and global-level network statistics (Fortunato, 2010; Newman and Girvan, 2004; Wasserman and Faust, 1994). Often, innovation researchers are interested in the composition of, and interaction between, close-knit groups of meso-scale objects such as groupings of similar technologies (Alstott et al., 2017; Balland and Rigby, 2017; Lee et al., 2015; Mejia and Kajikawa, 2020; Yan and Luo, 2017), and the application of community detection to the multilayer citation network leaves room for direct comparison between our results and these objects that we usually work with. Lastly, community detection can be applied to both multilayer and single-layer networks, which will allow for comparisons between the resultant communities.

2.3. Data

The multilayer citation network we construct is generated by citations made by triadic patents and only includes those made to and by triadic families. For the purposes of the current work, *triadic patents* are patents granted by one of the triadic offices that have family members, or equivalents, granted by the other two triadic offices. *Triadic families*,

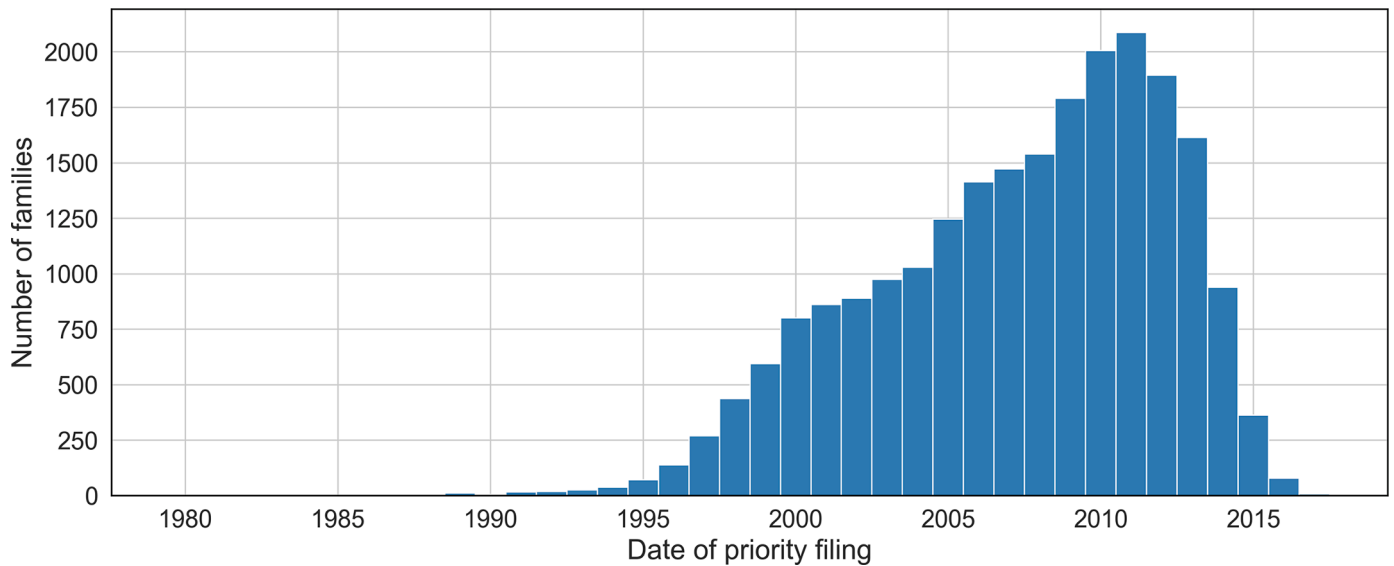


Fig. 2. Family priority dates. A histogram of the priority dates of the triadic families considered in this work, subject to the restrictions laid out in Section 2.3. All families have their US member granted in 2001 or later, but the earliest filing date can be considerably earlier.

on the other hand, will refer to the full set of documents belonging to a family that contains triadic patents.¹² These sets include both applications and patents and may be filed at or granted by offices outside the three triadic offices (provided that they are within a family containing triadic patents).

There are several reasons for choosing this subset of nodes and links to define our network, beyond the aforementioned desire to be conservative in our empirical design. The first is that we require well-defined layers. By restricting the citing patents to those granted by the triadic offices, the links (and, therefore, network layers) are defined by the citation context (e.g., the jurisdiction where it was made and the reason it was added), which isn't available for many offices. Second, restricting the cited families to those that are also triadic means that there are no cross-layer citations, which significantly simplifies the network from a mathematical perspective. For example, a US triadic patent citing a pre-grant publication that was only filed at the Japan Patent Office would be a cross-layer citation, as the latter node does not exist in the US layer. If, however, this Japanese publication was part of a triadic patent family, we can 'redirect' this US-originating citation to the US-granted family member, as this patent covers the same technical content, and the citation can remain within the US citation network layer where it was generated. Third, all triadic offices provide detailed citation data. There is no theoretical reason why citation network layers associated with other countries cannot be added if the data exists, but we deemed the triadic offices to be the best starting point to demonstrate the use of the multilayer framework due to their existing popularity among both applicants and researchers.

In this work, we also wish to demonstrate the importance of citation source and context. During the application and examination process, citations that reach the front page of the patent may be added by one of

¹² Note that this definition is slightly different to that used in previous work, notably [Dernis and Khan \(2004\)](#). Until the year 2000, applications to the USPTO were not published, so it was generally impossible to know whether equivalents were filed in all jurisdictions. This led to a slightly awkward definition (families with equivalents granted by USPTO and applied to EPO and JPO) that was in wide use until sufficient time had passed for USPTO application data to accumulate. A common definition in use currently is those families with equivalents filed at the triadic offices; however, as US applications do not list citations, we restrict this definition further to require a grant at each office.

several parties for a variety of reasons. One problem inherent in this citation metadata is that different offices have different examination guidelines and legal frameworks that inform how prior art is cited ([Higham and Yoshioka-Kobayashi, 2022](#)). Further, the way that these differences manifest themselves in the metadata that researchers can access is not consistent across offices or, indeed, across time. For some of the analyses in this work, we broadly group citations at each office into two groups: those that were likely found by the examiner and those that were likely found by the applicant. While these groups are far from perfect,¹³ we do so to illustrate the flexibility of the multilayer network approach—the citations that comprise each layer can be filtered based on the research purpose. This flexibility is discussed in more detail in Section 4. One minor restriction that accompanies this approach is that we require citation metadata to exist for all citing patents. The USPTO only started to include this metadata for granted patents from the start of 2001, so the triadic families we consider in this work are those for which the first US grant was in 2001 or later. All families considered in this work have all of their triadic members granted before April 2020. A histogram of the priority dates of the families that comprise the networks we consider in this work is displayed in [Fig. 2](#).

Most of the data used in this work were obtained from Google Patent Public Datasets.¹⁴ However, noting that, at the time of data collection, that data was not complete for citations between Japanese publications (notably, Japanese patents citing published Japanese applications), this data was supplemented by data supplied by the Intellectual Property Institute's Patent Database.¹⁵ We also make use of Cooperative Patent Classifications (CPCs); for consistency, we assign each family the classifications associated with their first US member, as determined by the USPTO. This data was obtained from PatentsView.¹⁶

To reduce the computational complexity associated with large networks, we prefer to work with a subset of the whole patent family network that nonetheless resembles the structure of the whole. Using a set of obviously technologically related families such as those in a specific technology class or filed by firms in a specific sector may not satisfy

¹³ This is particularly true for the JPO. However, there is suggestive evidence that applicant citations are more likely to be background art than art that could lead to a rejection of the application ([Okada et al., 2018](#)).

¹⁴ <https://tinyurl.com/googlepatentdata> (accessed 25/10/2021).

¹⁵ www.iip.or.jp/e/patentdb/index.html (accessed 25/10/2021).

¹⁶ <https://patentsview.org/> (accessed 25/10/2021).

Table 1

Layer descriptions. Descriptions of the layers considered in this work, alongside their abbreviations and the number of links found within them. All layers contain 22653 nodes, and there are a total of 63916 citations in the multilayer network (MULTI) that is comprised of the layers described in the first seven rows. The last two rows are single-layer networks obtained by flattening the two USPTO layers (US-AGG) and all seven layers (ALL-AGG), respectively.

Layer	Citing party	Abbreviation	Description	Links
USPTO	Examiner	US-EXM	Cited by examiner during patent prosecution	15607
USPTO	Applicant	US-APP	Cited by applicant through an Information Disclosure Statement and unused by examiner	23145
EPO	Applicant	EP-APP	Cited by applicant, in the patent text or otherwise	4326
EPO	Examiner	EP-ISR	Cited by examiner in an international search report	5732
EPO	Examiner	EP-SEA	Cited by examiner in an EPO search report	5206
JPO	Examiner	JP-REJ	Cited by examiner as justification for application rejection	4612
JPO	Examiner	JP-BCK	Cited by examiner as background information	5288
USPTO	All	US-AGG	Cited by anyone (USPTO patents)	38752
All	All	ALL-AGG	Cited by anyone (all triadic patents)	63916

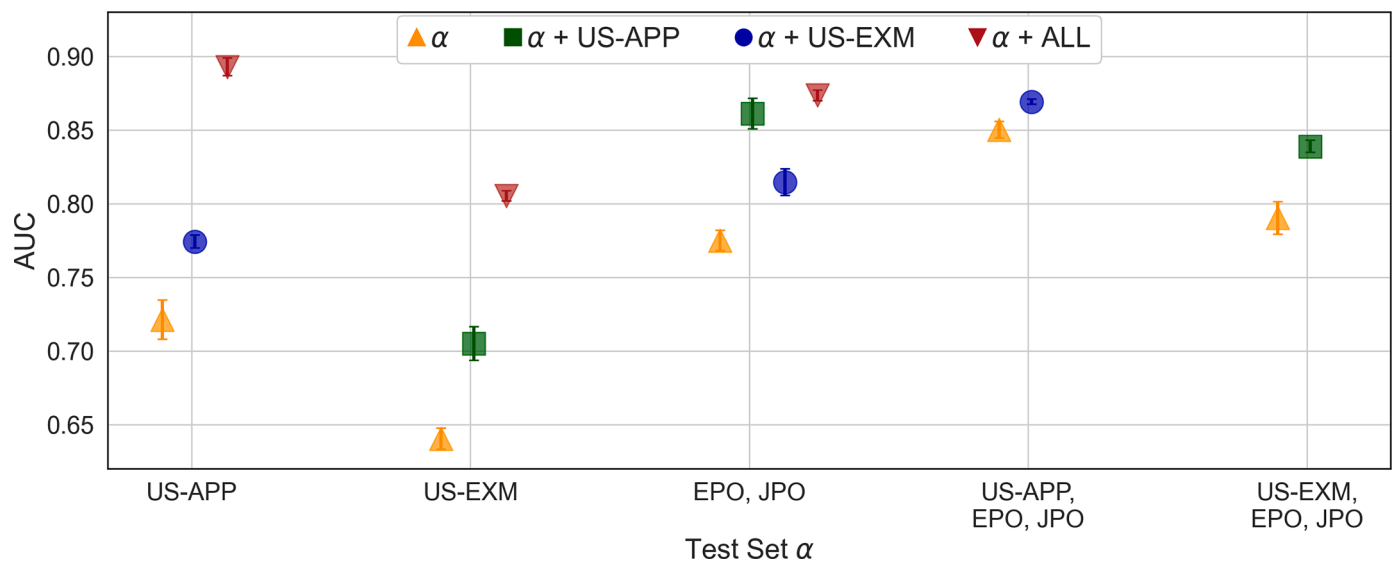


Fig. 3. Layer interdependence. The x-axis presents different target layers α , and the y-axis shows the AUC obtained through 5-fold cross-validation for measuring layer interdependence. Orange results refer to the baseline AUC, where the algorithm is only given access to that target layer. Green and blue markers show the increase in the AUC for the α set when the algorithm is given access to the US-APP or the US-EXM, respectively. The red points refer to the AUC obtained by giving access to all other layers in the network. The results displayed are averages and standard deviations over the 5 folds.

this requirement, given the known differences in citation patterns across fields (Alcácer et al., 2009; Higham et al., 2017). To remedy this, we choose the subset of patents assigned to CPC class Y02: “technologies or applications for mitigation or adaptation against climate change.” The Y02 class is always a secondary classification and can be added to patent families from a broad set of technologies, from those aimed at reducing drag on airplanes to those aimed at treating diseases whose impact may be exacerbated by climate change (Hašič and Migotto, 2015; Veeffkind et al., 2012). This class (and its subclasses) are commonly used as filters to study patented technological developments within specific domains related to both the mitigation of climate change, such as cleaner transport (Aghion et al., 2016; Barbieri, 2016) and energy production (Persoon et al., 2020; Sun et al., 2021), and our adaptation to the inevitable and wide-ranging environmental challenges we will face in the near future (Dechezleprêtre et al., 2020; Hötte et al., 2021). As such, we believe this technology class comprises a suitable microcosm within which we can effectively demonstrate the application of multilayer network methods to patent citation networks.

The resulting data set consists of a well-defined set of citing families, their CPC classifications, the citations they make,¹⁷ and the jurisdiction and context of each citation. A description of the layers considered in this work (which can be aggregated for specific empirical tests) can be

¹⁷ We exclude very rare citation types, such as those originating from third parties.

found in Table 1.

3. Methods and Results

3.1. Interdependence

Before a detailed examination into the kind of information that may be extracted from the multilayer network that is not accessible when using a single layer, it is first important to assess whether there is new information in the multilayer network at all. That is, if there is a high level of redundancy between the information contained in each network layer, then the case for using a multilayer framework is weakened. At the same time, if the layers contain very different structural patterns, then a multilayer framework may not be ideal, and more informative results may be obtained if they are treated as individual single-layer networks instead.

One way of assessing these properties is by measuring the interdependence of each layer, or set of layers, relative to the information that can be found elsewhere in the network. Several measures of interdependence have been proposed in the past (Morris and Barthelemy, 2012; Nicosia et al., 2013; Parshani et al., 2011), many of which take a random walk approach to the level of layer interdependence or ‘coupling’ of layers in the network. In this work, at a high level, we are instead interested in the degree to which the information contained in one network layer can inform us about the information contained in another layer.

To this end, we employ the method introduced by De Bacco et al. (2017) and described in detail for our case in Appendix A.2. This method is a link prediction exercise, whereby a randomly-selected portion of the target layer or layers α have their link information removed and the remaining information in the network may be used to predict the existence of links. As a baseline, the remaining portion of the α is used as the training set, the receiver operating characteristic (ROC) curve is calculated, and the area under this curve (AUC) is computed. We can then introduce sets of other layers, β , into the training set, and compare results obtained by adding this information to those of the baseline. If the predictive power (as measured by the AUC) of this augmented set $\alpha + \beta$ is not significantly larger or smaller than the baseline predictive power, then β does not contain useful information over and above that contained in α . If, however, we note a significant increase in predictive power relative to the baseline, then β contains complementary information that cannot be extracted from what remains of α .

Much information can be garnered from comparisons of the change in predictive power when α and β are interchanged. For example, when the links in one layer are a subset of links in another, then we expect the change in predictive power to be asymmetric when we swap α and β — adding the subset to try to predict links in the full set will likely produce worse results than if only the full set was used for training the model. The information in the subset is redundant and could even mislead the model.

When two layers contain complementary information we would expect increases in predictive power regardless of the layer comprising the test set. This complementarity can arise in several ways, such as through similar community structure despite large differences in the specific links that produce these structures. A significance decrease, on the other hand, would indicate that β contains information that is irrelevant for the prediction task and actually added noise; this could occur, for example, if the link generation mechanisms were independent of the node properties, or were driven by different node properties in different layers.

Fig. 3 shows the results of the interdependence analysis for various α and β sets in which we are interested. For graphical simplicity, we focus on the sublayers generated by the USPTO and the complete JPO and EPO layers (where the latter two always include all of their sublayers listed in Table 1). This is done to demonstrate, compactly, the complementarity of information across jurisdictions as well as that of their sublayers, with the most commonly utilised sublayers in the literature (US applicant and examiner citations) as exemplars for the latter calculations.

The results displayed in Fig. 3 show that adding more layers increases predictive power across all combinations of α and β we considered. This outcome suggests that, while they differ by the amount of unique complementary information they contain, each layer nonetheless contains information that is not available in the other layers. Specifically, information about the missing values in α is more accurately predicted when layers that are not already in α are included in the training set, relative to the sole use of the information that remains in α . This is to be expected, as examiners at each office conduct much of their prior art search independently.

A prime example of complementarity is displayed by the US sublayers (US-APP and US-EXM). These layers are almost mutually exclusive,¹⁸ but predictive power for links in one layer is significantly boosted when the other layer is added, regardless of which is the test set. That is, there is very little overlap in these layers, and yet one can be successfully used to predict the links in the other, likely due to the similarity of mesoscale network communities within each of these layers.

That some citation types add more information than others is also expected. After all, information sharing occurs regularly between offices (Wada, 2020), and this process leads to the duplication of citations

between specific layers. While this sharing happens increasingly through direct collaboration between offices examining equivalents,¹⁹ most of the citations we consider here were made before these formal programs were launched. As such, for much of the time period we consider, the information ‘sharing’ likely takes place indirectly, through applicants. For example, the EPO produces a search report for the applicant to consider before a substantial examination takes place. Under their duty of disclosure obligations at the USPTO, it is considered good practice to pass this information on to the USPTO if an equivalent is being examined there simultaneously (which will usually be the case for triadic patent families). This information is submitted via an information disclosure statement and the USPTO examiner then assesses the relevancy of the prior art that is listed on the search report. When it happens at all, only a small percentage of citations from the EPO search report will be used to justify rejection and be recorded as examiner citations, while the remainder will be recorded as applicant citations. As such, the EPO search report is a non-obvious mechanism through which citations are duplicated from EP-SEA citations to US-APP citations (and sometimes to US-EXM citations).

Similarly, while there is a knowledge disclosure obligation at the JPO, the incentives for complying are very weak relative to the USPTO (Nakamura and Sasaki, 2016). However, applicants to the JPO often use in-text citations to make a case for patentability, and perhaps much more so than the typical applicant to the USPTO or EPO. As such, it is plausible that, for triadic patents, these citations are included in-text in other equivalent applications and are therefore easily accessible to examiners in all jurisdictions. If these citations are deemed relevant by multiple examiners, these citations might also appear to be duplicated across network layers.

3.2. Community detection

Having found that the different layers likely contain complementary information, we now investigate the patterns extracted from a multi-layer network approach and compare them with those extracted from single-layer networks that exclude citation context. Specifically, we wish to detect communities of triadic families that are similar in their citation patterns. These communities represent mesoscopic structural patterns contained in the networks that are not objectively or directly observed, but can be inferred from the data.

To this end, we apply a community detection algorithm to three networks: i) the (seven-layer) multilayer network containing the EPO, JPO, and USPTO layers (MULTI); ii) the network obtained by flattening all the layers in (i) into a weighted single-layer network and ignoring citation origin and context (ALL-AGG); iii) the weighted single-layer network obtained by flattening the USPTO examiner and applicant layers only (US-AGG). Each link is weighted by the sum of link weights across all layers we consider; that is, if the same family-family citation exists once in each of n layers, then the link is assigned weight n . While rare, link weights greater than one can occur within sublayers; for example, when a divisional makes the same type of family-family citation as its parent, the link weight corresponding to this link will be two.

To perform the community detection task, we consider a probabilistic generative model that assigns a probability to a citation between two families that depends on the communities they belong to, as described in De Bacco et al. (2017). In our case we have access to relevant metadata about each triadic family, hence we consider the model of Contisciani et al. (2020), MTCOV, that is also able to incorporate the office at which priority was filed (which is often *not* a triadic office) as a node covariate to drive inference along with the network structural information. This covariate allows us to incorporate the home-bias of citations in early search reports (Bacchiocchi and Montobbio, 2010) and, to a lesser extent, industrial agglomeration patterns

¹⁸ Copying occasionally happens due to the recycling of citations for continuing patent applications.

¹⁹ <https://www.wipo.int/case/en/>.

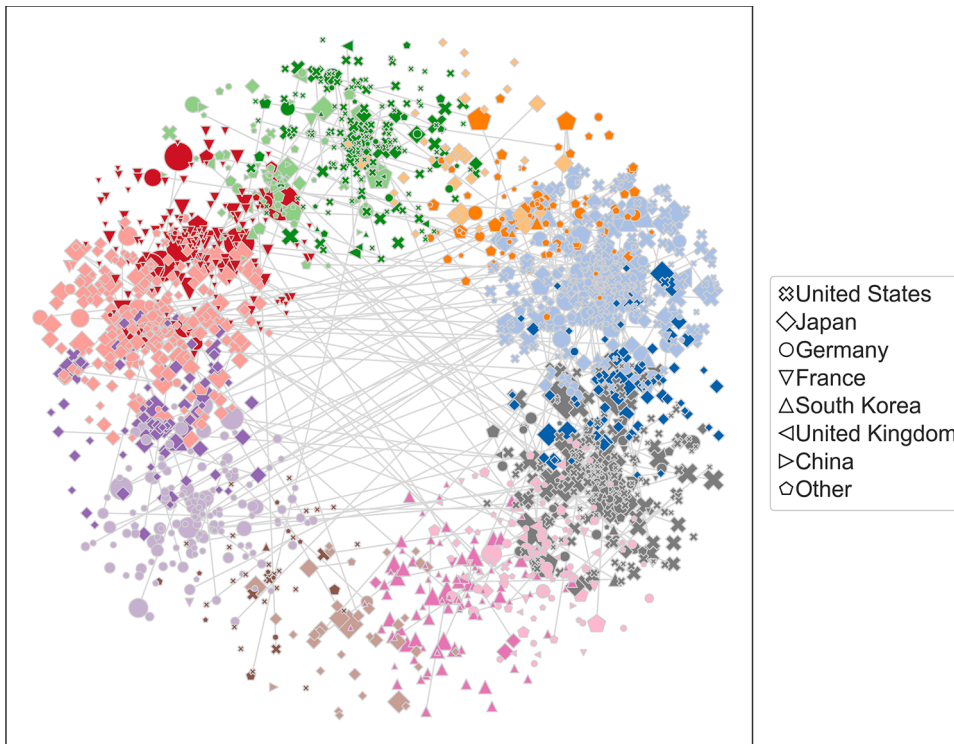


Fig. 4. Community extraction. This diagram shows the hard community membership partitions for MULTI. While inference was performed on the whole network, here we use a random sample of 2000 nodes and include any incidental links among these, for graphical clarity. The colouring shows the 15 communities found within the MULTI network. Node size is proportional to the number of outgoing and incoming citations, while node shapes denote the location of the assignee of each patent family.

(Asheim and Gertler, 2005) (given the strong correlation between assignee location and priority office), to inform the inferred citation probability alongside explicit network structure. This model automatically balances the weight of the covariates' contribution in determining the communities. In all our experiments we find that node covariates are indeed significant, in that they allow us to better quantify the probability of certain citation patterns. The optimal number of communities in each case is extracted through a cross-validation procedure, see Appendix A for details. In addition to being able to incorporate a covariate that may inform network structure at scales beyond individual links, MTCOV is scalable to large networks, allows overlapping communities, and is open-source,²⁰ all of which are desirable features for the current work.

We chose ALL-AGG as a comparison because it contains all the same links as the multilayer network, and even accounts for link overlap among layers, but without context. As such, any differences in the extracted communities arise solely due to the addition of citation context, and the incorporation of this context into our network model. US-AGG is included in these comparisons as the most common citation network used in previous work. The USPTO also tends to make many more citations per patent, and so this single-jurisdiction layer is likely to be the most 'complete', with respect to the links in the full triadic network.

The communities extracted for MULTI are shown for a random subset of patent families in Fig. 4. Analogous figures for the ALL-AGG and US-AGG networks can be found in Appendix B. While the model allows for overlapping communities (nodes can belong to multiple communities), in Fig. 4 we colour nodes by their 'hard' communities, whereby each patent family is assigned to the community to which it displays the highest affinity. The optimal number of communities, calculated via the cross-validation exercise described in Appendix A.1, was found to be 15 for the multilayer citation network and 7 for ALL-AGG and US-AGG. Finally, the location of the assignee of each patent family (rather than

the priority office, which is used as a covariate in the community detection procedure) is indicated by the shape of the node.

Between networks, several graphical observations can be made in the geographic composition of the extracted communities, despite the differing community sizes. First, country-based homophily is very clear. The most obvious example of this is that families filed by Japan-based assignees are primarily grouped with families that are also filed by Japan-based assignees, with the only observable difference between the networks being how many communities are found within this group of families (1 for ALL-AGG and US-AGG, and 5 for MULTI); however, this difference is expected as the optimal number of communities multilayer network is greater. The other consistently geographically-homogeneous communities include those families assigned to German firms and those assigned to South Korean firms. The existence of these groupings is

Table 2

Network comparison and diversity measures. Here we show the results of the network distance calculations as well as the diversity measures. DeltaCon (DC) and Frobenius (F) distances between the class-projected networks (leftmost block) and the externally defined co-classification (Co-class) and inter-class citation linkage (IC Cites) networks are displayed in the central block. The median Rao-Stirling class diversity (RSD) across communities and the median Herfindahl-Hirschman Index (HHI) of classes' dispersion across the C communities are shown in the rightmost block. * indicates non-optimal partition. The lowest values within each comparison set are highlighted in bold.

	C	Comparison Network				Diversity	
		Co-class		IC Cites		RSD	HHI
		DC	F	DC	F		
MULTI*	7	32.43	115.98	30.40	115.01	6.81	0.185
ALL-AGG	7	34.00	149.40	31.96	148.31	7.02	0.263
US-AGG	7	34.54	150.63	32.49	149.59	7.05	0.271
MULTI	15	30.08	58.66	28.09	58.61	6.87	0.131
ALL-AGG*	15	31.47	78.87	29.49	78.56	6.87	0.169
US-AGG*	15	30.79	76.73	28.81	76.42	6.99	0.167

²⁰ <https://github.com/mcontisc/MTCOV>

somewhat expected — geographic citation biases are a well-known phenomenon and have a wide range of drivers, including local industry agglomeration, shared language, prior-art search strategies, knowledge spillovers, and coordinated technological development strategies at the national level (Almeida and Kogut, 1999; Bacchiocchi and Montobbio, 2010; Jaffe et al., 1993; MacGarvie, 2005; Wada, 2016). Because priority office information is included in the community detection algorithm, the existence of the kind of geographic grouping we observe reflects that while technological similarity plays a big role in citation linkage at the micro-level, simple geographical metadata can be highly predictive of network structure at larger scales.

3.3. Network Communities and Technological Similarity

One would expect that the citations we consider in this work *should* link families with technological similarities and, therefore, the communities detected should group inventions with shared and legally relevant technological features. Indeed, the geographical biases in citation linkages that are observed above may be considered to be artifacts of the systems within which technological development occurs, and perhaps even hinder our understanding of the nature of innovation more generally. We assert that the multilayer framework is one way of mitigating some of these biases, as it integrates relevant technological relationships uncovered by several different, and geographically separated, patent agents and examiners working mostly independently. In aggregate, this information should give a more balanced view of technological similarity and down-weight those links that are heavily influenced by unwanted geographical and office-specific biases and conventions. However, the link weights in the ALL-AGG network may play a similar role. As such, we will now turn to the differences in the technological information contained in the three networks and examine the importance of citation context (i.e., source and justification) in assessments of technological similarity.

To do this, we directly compare the network of meso-level technological relationships that can be gleaned from extracted communities with externally-defined technological categories. First, we construct a weighted bipartite (two-mode) network of relationships between the extracted communities and the 3-digit Cooperative Patent Classification (CPC) codes that the families within each community were assigned upon application to the USPTO.²¹ CPC codes, henceforth referred to simply as *classes*, were chosen due to their status as the primary classification system at two of the triadic offices and widespread use in research, particularly in studies of technological evolution and forecasting technical change. The weight of each link in the bipartite network between communities and classes is proportional to the fraction of families in each community that were assigned to a given class. We then project onto the technology class nodes to obtain a network of classes wherein links exist between classes that were both found in the same community or communities. A higher link weight between two nodes in this projected network reflects a more similar distribution of those classes across the extracted communities (Vasques Filho and O’Neale, 2018).²²

We then construct two basic comparison networks: co-classification (Breschi et al., 2003; Engelsman and van Raan, 1994) and inter-class citation linkage (Alstott et al., 2017; Leten et al., 2007). The former

²¹ This choice was made for the sake of consistency. Different offices may make slightly different judgements regarding the particular set of classes assigned to an application. By using data from a single office, we do not have to be concerned with these systematic differences.

²² Note that these classes are not directly used in the community detection process. However, the community detection process relies on citation linkages, and these citations are often found through searches within the technology classes to which the application under examination has been assigned (see, e.g., Demey and Golzio, 2020).

contains a link between (3-digit CPC) classes when a family is assigned to both, with weights proportional to the relative frequency of such occurrences. The latter network contains links between these classes with weights proportional to the number of citations made between families that were assigned to each class, normalised to the total made by each of the classes.²³ We keep self-loops in this network, as they are required for sensible link-weight normalisation. For example, if class A makes 10 citations (and receives none), one of which goes to a class B family but 9 return to other class A families, this is a very different situation from one in which all 10 go to class B families. Because we normalise link weights by total citations made, ignoring self-citations would give the link from A to B the same weight in both scenarios, rather differing by a factor of 10. Further description of the construction of all networks used in this section can be found in Appendix B.

Now that we have two externally defined, node-aligned class networks, we are able to directly compare their structure to those extracted from the community-class bipartite networks. Because the nodes in each network we wish to compare are labelled and the same for all networks, we are able to use known-node correspondence methods that allow for comparisons at the node-level in such a way that accounts for differences in relationships between specific node pairs and for higher-order relationships (Tantardini et al., 2019). For this exercise, we use two different methods of comparison: the Frobenius norm and DeltaCon (Koutra et al., 2013).

The Frobenius norm is applied to the raw differences in the adjacency matrices between two networks, and thus quantifies the entry-wise (link-level) differences in the matrices being compared. When the networks being compared are unweighted, this distance is simply the square root of the number of pair-wise differences between the networks. However, this method easily accommodates the weighted case, wherein each pair-wise difference can have a magnitude other than unity.²⁴ The Frobenius norm is a crude comparison method that cannot account for higher-order relationships between nodes, such as the importance of a link in the overall structure of the network, but it is a good heuristic when making multiple comparisons as we do here. DeltaCon, on the other hand, is more sophisticated, and indirectly considers every possible path between two nodes. In this way, differences in the weights of links that are particularly important for the network structure at the macro-level are incorporated into the comparison. While the DeltaCon algorithm can be very computationally expensive on large networks, and an approximation is possible, our class network is small enough (535 nodes) that the exact form can be used (Koutra et al., 2013). Both the Frobenius norm and DeltaCon calculate a distance metric whereby smaller distances indicate more similar networks. These methods are implemented in Python using the numpy (Oliphant, 2006) and netrd packages (McCabe et al., 2021).

In addition to the network comparison methods, we are also able to quantify the diversity of technology classes within each community extracted. For this purpose, we make use of the Rao-Stirling diversity (RSD) (Rao, 1982; Stirling, 2007), which considers both the homogeneity of each community (with respect to the classes within it) and the level of ‘surprise’ that specific pairs of classes are found together. For the latter consideration, we operationalise class distance using the

²³ To compare this network to our (undirected) projected networks, we take the sum of the normalised weights of the directed links between classes to obtain an undirected link weight. This simplification is a necessary evil for the current purpose and may miss some nuance in certain technological relationships.

²⁴ Specifically, the Frobenius norm of a matrix $A_{m \times n}$ is defined as the square root of the sum of the absolute squares of its elements,

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}. \quad (1)$$

inter-class citation network described above, as citations are what we use to extract the communities in the first place.²⁵ Calculating this index for all communities extracted from a particular network, we take the median index across these communities as a measure of their average diversity. The RSD is high for a particular community when classes co-occur in high proportions with other classes with which not many citations are exchanged. RSD is low when classes generally only co-occur in high proportions with other classes with which they exchange many citations. Specifics of the RSD can be found in [Appendix B.2](#). This kind of analysis, when compared to the network comparison methods above, may be considered relatively myopic. It can only capture the internal composition of individual communities without accounting for the relationships between the pairs of technologies in other communities. However, this calculation may provide insight into the origin of differences we find for the network-level comparisons.

Lastly, we calculate the spread of technologies across the extracted communities. A priori, we do not know what the relationships between the communities are, so we cannot integrate a distance metric to account for the level of surprise that a family assigned a particular class is found in a given pair of communities (as we did for the previous diversity measure). As such, we implement the Herfindahl-Hirschman Index²⁶ (HHI) (Herfindahl, 1950; Hirschman, 1945; 1964; Simpson, 1949) to measure the extent to which classes are splintered across communities. The details of this calculation can be also found in [Appendix B.2](#). The HHI is maximised when all families assigned a particular class are in the same community and minimised when there are the same number of these families in each community. Again, the median HHI across all communities is compared across the networks. Like the Rao-Stirling index above, this calculation may add additional colour to the more comprehensive network comparisons. It is important to note that while we believe that it is desirable that communities are able to capture, to some extent, the large-scale structure of the technology-level networks, neither the spread of technologies across communities, nor the internal diversity of communities, is a test of the performance of the community extraction exercise.

It is important to note that the optimal parameters for the community partitions for the three networks are different—the optimal number of communities found for the multilayer network is 15, while for the others it is 7. For this reason, we run the community-detection algorithm for each of the non-optimal partitions (7 for the multilayer network and 15 for the others) to obtain a complete set of networks with which we can make fair comparisons. In sum, we construct six bipartite (community-class) networks which we project onto the class nodes to compare with the co-classification and inter-class citation networks.

The results of this analysis can be found in [Table 2](#). First, we find that the communities in the multilayer network generate class networks that are more similar to the co-classification and inter-class citation networks than those generated by the other two networks. This finding holds for both individual-link-level comparisons (Frobenius) and when higher-order relationships are taken into account (DeltaCon), for both optimal and non-optimal partitions of the multilayer network. Further, we find that the average RSD of the individual communities is lowest, while classes are the most evenly distributed across the communities (low HHI), in the multilayer case.

These observations lend themselves to some interesting interpretations. When looking at all communities, in combination, those extracted from the multilayer network imply technological relationships that are closer to the explicit technology networks than the flattened or single-jurisdiction approaches. However, the diversity calculations suggest that this observation is not simply driven by the extraction of homogeneous communities that group technologies in a straightforward

manner. In fact, technology classes are more thinly spread across communities in the multilayer case, while the average internal diversity of classes is generally lowest for this network once known technological similarities between classes are accounted for. This suggests that, on the micro-level, the multilayer (relative to the single-layer) network approach is more sensitive to citation linkages than co-classification, but is nonetheless better able to represent real technological relationships on the meso- and macro-levels.

Indeed, our results are consistent with the conclusion that the multifaceted nature of the technological relationships that are embedded in citation data may be partially lost when a multilayer network is flattened into a single-layer one. This view rests on an assumption that different technology types can be related to each other in different ways. For example, let's assume that applicants filing a patent assigned to class A prefer to cite families assigned to class B, while examiners examining the same patent prefer to cite those assigned to class C. When, such as in this example, these different relationships are driven by different citing parties, the erasure of citation context will lead to the loss of this nuance. This problem may be exacerbated in the presence of higher-order effects, such as if the above citation behaviour only occurs when a fourth class D is also assigned to the patent application. In contrast, the multilayer network approach ensures these nuances and higher-order relationships remain accessible. The retention of this kind of technologically relevant information, particularly with respect to rare or subtle inter-class relationships, would be consistent with the findings displayed in [Table 2](#).

4. Discussion and Conclusions

Historically, research informed by patent citation data has often ignored citation source and context. There can be a perfectly reasonable reason for this practice, such as when one is only interested in citations made to and from patents in a single jurisdiction to study, for example, the effect of a local policy change. However, a truly comprehensive and global view of patented inventions and the relationships between them is only possible when data from multiple sources are integrated sensibly. It is in these contexts that the multilayer network is a natural framework for analysis.

In this work, we introduce the concept of multilayer patent citation networks as a natural way to present and analyse global patent information without loss of citation context. We conduct several empirical analyses to demonstrate the utility of the multilayer framework. All analyses are conducted on a subset of the full citation network, containing all triadic patent families classified into CPC class Y02 with US members granted from the year 2001. By design, this subset will give the most conservative estimates of the additional information that may be extracted from the multilayer network relative to its single-layer counterparts. Our results in this work suggest that not only is there, indeed, a considerable amount of additional information contained in the multilayer citation network relative to those single-layer counterparts, but this information is technologically relevant and captures nuanced aspects of the technological relationships between patented inventions.

First, an interdependence analysis shows that additional network layers, defined by citing office, contain complementary (rather than redundant) information that may be used to predict the link-level structure of other layers. To test whether this complementary information is important for characterising network structure more generally, we then conduct an exercise in community detection. This is carried out and compared across three different networks: the multilayer network, the flattened and weighted (single-layer) version of the multilayer network (containing all the links in the latter but without citation context), and the complete (flattened, single-layer) US citation network that is most commonly used in technological network analyses. While there is a notable similarity in the communities extracted from these networks, there is also significant disagreement, indicating that the information contained in the citation context may be important for

²⁵ That is, it would be a 'surprise' to find a pair of classes that don't cite each other, but are nonetheless found in the same community.

²⁶ Sometimes referred to as the Simpson index.

characterising the mesoscopic structure of the global citation network.

To test whether the differences in community structure are technologically meaningful, we conduct direct comparisons between the technological relationships implied by the extracted communities and those of previously studied meso-scale networks of technological similarity: the co-classification and inter-class citations networks, at the CPC 3-digit level. These tests are conducted, in part, to show how the information content (i.e., citation context) contained in citation networks can be related to the meso-scale technological structures that are perhaps more established in the technology management community. To be able to draw a direct comparison, we construct the bipartite networks between communities and classes, then project onto the class nodes to obtain a class network wherein links reflect levels of co-occurrence in the communities. To add colour to these comparisons, we also compute the Rao-Stirling diversities of these communities (across classes) and the Herfindahl-Hirschman Indices of class (across communities). Relative to the flattened networks, we find that while the communities extracted from the multilayer network are less diverse and the implied class network more similar to the co-classification and inter-class citation networks, classes are more evenly spread across communities. These results suggest that citation context is technologically relevant and a more realistic mesoscopic network structure can be inferred when we depart from the view that technological relationships are mono-faceted or driven by simple class-level technological similarity.

While we include the US citation network in our comparison exercises, this is only done as an acknowledgement of its position as the dominant data source in the extant literature. The flattened version of the multilayer network, on the other hand, contains all the links that are present in the multilayer network, but without the context that allows us to define the layers. As such, we consider this network the most appropriate comparison network, as any differences found must be driven by the absence of citation context. That the communities extracted from the multilayer network more closely replicate the established and explicit co-classification and inter-class citation networks indicates that citation context adds technologically relevant information in the aggregate, despite displaying higher within-community diversity of classes. This suggests that ignoring citation context results in a bias towards within-class citations (that are easier for all parties to search for and find), at the expense of the rarer inter-class citations and class combinations that play a larger role in both the network structure as a whole and, arguably, technological progress in the long-term (Castaldi et al., 2015; Kelly et al., 2021; Mewes, 2019; Verhoeven et al., 2016). Considering citation generation mechanisms, it is plausible that citation context provides important clues as to the relevance and nature of the technological relationship between citing and cited inventions (Alcácer et al., 2009; Azagra-Caro et al., 2011; Criscuolo and Verspagen, 2008; Kuhn et al., 2020; Li et al., 2014). As such, treating all these links as equal, with respect to their information content, is clearly not ideal for many use-cases.

4.1. Limitations

The main limitations of the empirical analyses conducted in this work are those restrictions we placed on the families we chose to include. As we describe in Section 2.3, these restrictions were put in place for a variety of reasons, including data availability, computational limitations, and a desire to demonstrate our approach in a conservative manner. Little can be done about data availability; however, this only affects our ability to examine citation context in the US case, and only for times earlier than the year 2001. In any case, we suggest that families granted after this time provide a sufficiently large sample for the purposes of this work.

The conservativeness of our approach is introduced with the decision to consider only those families with granted patents at all three triadic offices. This means that all offices had access to the same set of prior art,

and had the opportunity to share information among themselves. In turn, this would introduce maximum redundancy between layers, and minimise the additional information that can be added by the inclusion of citation context. It is for this reason that we think of our approach as conservative. Extensions of the restrictive, special-case multilayer framework that we examine here are discussed below in Section 4.2, and highlight the potential of this framework going forward.

Lastly, to reduce the computational complexity of our analyses, we restrict the included families to those classified into CPC class Y02. While we maintain that this subset is an appropriate representation of the patent citation network as a whole, there may be arguments against its generalisability. However, in the case that this class contains a more homogeneous set of families than the set of all families (which is almost certainly true), then the inter-class structure that we are able to explore is likely to be *less rich* and *less nuanced* than that of the full network. Detecting higher-order nuances is precisely the domain in which we suggest the multilayer network excels, so following this logic would lead us to conclude that the current approach is, again, a very conservative one.

4.2. Future Work

This work aims to describe the construction of multilayer patent citation networks then conceptually and empirically justify their use. This framework may prove to be of particular interest to those who would prefer representations of technological relationships that are not as sensitive as extant frameworks to the idiosyncrasies of individual patent offices. However, both the layers that are selected to comprise the network and the appropriate empirical methods to extract information from this network will depend on the specific use-case. Here, we describe the myriad methodological doors that are opened with the introduction of patent-based multilayer networks into the broad field of science, technological, and innovation studies.

The obvious extension to the current work is to take a less conservative approach with respect to the subset of families and citations considered. This can take the form of additional layers, nodes, or links. The addition of layers corresponds to the addition of new citation contexts (such as in-text citations (Verluisse et al., 2020)) or the addition of new jurisdictions. The addition of nodes and links, on the other hand, would relax the condition that a family be triadic. Citations between triadic families only make up a tiny portion of all citations made and received by these families. For example, in Fig. 1, we show the triadic ego network of the family with USPTO equivalent US-6819081-B2. In this restricted network, this family only receives 4 citations from other triadic families classified into class Y02. If we remove all restrictions on the patents we include in our network, however, this family receives almost 50 citations; about 90% of these are from families that have a triadic member, and about 95% are from families that are also classified into Y02. As such, removing the triadic family requirement but keeping the network restricted to the triadic offices and the Y02 classification would dramatically increase the sample size.

Multilayer citation networks can also be flexibly aggregated. Just as one can analyse the inter-class citation network for a single jurisdiction or citation context (e.g., US applicant citations), it is also possible to include additional layers containing the equivalent information for other jurisdictions or contexts. In fact, in the same way that we use families to align layers in the current work, any metadata that connects groups of patents between network layers forms a natural multilayer configuration. Classes, firms, and inventors can all be linked across jurisdictions and citation contexts and their networks analysed in a multilayer framework. Even in the single-jurisdiction US case, for example, the relative positions of firms in the inter-firm citation network will depend on whether one uses examiner citations, applicant citations, third-party citations, or in-text citations. Because firms can be represented as nodes across all of these context-specific networks, multilayer network tools may be applied to obtain a comprehensive and integrative

view of the network structure without abandoning citation context. Citations to non-patent literature such as scientific articles is challenging to incorporate into patent citations networks generally, but it is certainly possible to treat this information as family-level metadata — perhaps to construct a bipartite network similarly to how technology classification was used in Section 3.3. More complicated uses of this information could match institution and inventor data from patents onto scientific articles to extend recent work on the multilayered interplay between authorship and the broader dynamics of science and collaboration into the technological domain (Nanumyan et al., 2020; Omodei et al., 2017; Zingg et al., 2020).

In addition to these data extensions, the conceptual arguments against the omission of citation context lead to a strong case for the further application of novel tools designed for the study of multilayered systems. To return to the public transport analogy, it would be unwise to treat all modes of transport as equal if you are trying to find the fastest route between two places in the network. In the same way that the time and financial costs of using different modes affects the route choice between two points in a physical landscape (which will be moderated by the amount of time or money you had), citation networks are embedded in a technological landscape (Fleming and Sorenson, 2001; Kauffman et al., 2000) and different types of citation may traverse this landscape in different ways. This intuition has significant consequences for the analysis of citation networks. For example, any algorithm that ‘walks’ through the network, such as PageRank, should consider the ‘cost’ of each link in a similar way to one plotting a route through a multilayered transportation network. The application of multilayer network methods opens the door to a menagerie of new analytical tools to develop more sophisticated and tailored metrics for studies of technical change and the nature of innovation systems. For example, the identification of patent thickets (Bessen, 2003; Shapiro, 2000) is often conducted through, or supported by, citation network analysis (Von Graevenitz et al., 2011; Yuan and Li, 2020; Zingg and Fischer, 2018). The multilayer framework may assist in these studies — thicket identification depends crucially on the citation context (blocking vs. non-blocking citations) and the jurisdiction (a thicket is necessarily a single-jurisdiction phenomenon). Adding citation context and linking families across jurisdictions for direct comparison may allow for thickets to be more easily distinguished from fields with dense, but non-overlapping, intellectual property rights. For example, when calculating clustering coefficients in multilayer networks, one can specify weights for different kinds of citation or penalise cycles that move between layers (De Domenico et al., 2013). This kind of flexibility can be used to operationalise the definition of thickets in a way that doesn’t simply ignore applicant-provided citations or citations from other jurisdictions, which may not be entirely irrelevant, particularly at the firm level.

Network centrality is another important concept that is generalised in the multilayer case (De Domenico et al., 2015; Solá et al., 2013; Solé-Ribalta et al., 2014; Taylor et al., 2021), and can also be readily applied to citation networks. For example, without citation context, it is hard to know whether firms are central because they block the patents of

competitors or are a source of knowledge from which other firms build. Further, firm centrality will likely depend on the jurisdiction one examines, so multilayer centrality may give a more holistic view of their centrality in global markets.

Both technology roadmaps (Lee et al., 2009) and technological trajectories (Verspagen, 2007) may be significantly altered by the incorporation of citation context, as different kinds of citation appear to hold different information, which may, in turn, be useful for forecasting or tracing different kinds of technical change (Acemoglu et al., 2016; Mariani et al., 2019). So-called ‘main paths’ in technological trajectory analysis (Hummon and Dereian, 1989; Verspagen, 2007) could be particularly sensitive to the weights that are placed on, or empirically determined for, different layers or citation contexts. The multilayer framework may also conceptually aid traditional economic analyses (Cai and Li, 2019), for which it is possible, for example, to allow layers to differ in importance when constructing proxy network variables that attempt to capture an abstract concept.

Lastly, pair-wise interactions may not be sufficient to describe the complex behaviour of interactions between the components of innovation systems that are accessible through citation networks. In particular, the interactions between firms or technology types that are visible in citation networks may be better represented through higher-order interactions (Battiston et al., 2021; 2020; Lambiotte et al., 2019). For example, the patenting and citing behaviour of firms may be described at several different scales. Higher-order representations allow us to differentiate changes in citation behaviour of a firm in response to sector-wide changes from the pairwise interactions between a firm and every other firm in its sector. Higher-order interactions can exist within layers of multilayer networks and it is possible that different higher-order behaviours are observable in different patent systems. In any case, it is clear that applications of network frameworks beyond single-layer networks with dyadic links are very much in their infancy in the field of innovation studies, and hold huge potential as more realistic abstractions of innovation systems.

CRediT authorship contribution statement

Kyle Higham: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Martina Contisciani:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing, Visualization. **Caterina De Bacco:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing, Visualization.

Acknowledgment

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani. MC and CDB were supported by the Cyber Valley Research Fund.

Appendix A. Model description

For the layer interdependence and community detection analysis we use MTCOV,²⁷ the model developed by Contisciani et al. (2020). MTCOV is a probabilistic generative model that incorporates both the topology of interactions and node attributes to extract overlapping communities in directed and undirected multilayer networks. It works also with single-layer networks, since this is the special case for which there is only one layer in the ‘multilayer’ network. The model assumes conditional independence between the network and attribute data, given a set of latent variables (including the node community memberships). The likelihood function is a linear combination of the network and attribute information, adjusted by a scaling hyperparameter $\gamma \in [0, 1]$, which controls the relative contribution of the two terms: for $\gamma = 0$ the model only considers the network topology, while for $\gamma = 1$ it only considers the attribute information.

²⁷ <https://github.com/mcontisc/MTCOV>

MTCOV has four parameters: two membership matrices accounting for outgoing and incoming links respectively, an affinity tensor that describes the density of links between each pair of groups among the different layers, and a parameter that matches communities and node attributes. The inference is performed with an Expectation-Maximization algorithm, and its implementation is efficient and scales to large datasets (such as the one studied here) because it exploits the sparsity of the dataset.

A1. Cross-validation and hyperparameter settings

MTCOV has two hyperparameters, the scaling parameter γ and the number of communities C . For each network under analysis, we estimate the hyperparameters by using 5-fold cross-validation along with a grid-search to range across their possible values. For the current work, we choose to vary $C \in \{2, 3, 5, 7, 10, 12, 15\}$ and $\gamma \in \{0, 0.3, 0.5, 0.7, 1\}$. Specifically, we divide the dataset into five equal-size groups (folds), selected uniformly at random, and give the models access to four groups (training data) to learn the parameters; this contains 80% of the matrix entries and covariates. One then predicts both links and node attributes in the held-out group (test set). By varying which group we use as the test set, we get five trials per realization. For performance metrics, we measure the area under the receiver-operator characteristic curve (AUC) (for the link prediction) and the accuracy (for the node attribute prediction) on the test data, and the final results are averages over the five folds. The AUC is the probability that a random true positive is ranked above a random true negative; thus the AUC is 1 for perfect prediction, and 0.5 for random chance. The accuracy classification score is 1 for perfect recovery and 0 in the worst case of overfitting. In order to choose the best pair of hyperparameters $(\hat{C}, \hat{\gamma})$ we look for the pair that performs best across both AUC and accuracy in the test set.

Since the networks are large, it is not always possible to compute the AUC on the whole training and test sets, hence we proceed with samples. In detail, we fix the number of comparisons we want to evaluate, here 10^5 , and for both the train and the test sets we sample 10^5 values from zeros entries (where there is no existing link) and we compute the link prediction on that sample (we save these values in a vector R_0); we do the same with the non-zeros entries (we save these values in a vector R_1). We then make element-wise comparisons and compute the AUC as:

$$AUC = \frac{\sum(R_1 > R_0) + 0.5 \sum(R_1 == R_0)}{|R_1|} \quad (2)$$

where $\sum(R_1 > R_0)$ stands for the number of times R_1 has a higher value than R_0 in the element-wise comparison; and $|R_1| = |R_0|$ is the length of the vector which is equal to the number of comparisons we fix. Moreover, when the network has a number of nodes bigger than 5000, we run the algorithm by computing the likelihood only on a batch of nodes (here a random subset with 5000 nodes) to speed up the computational time.

Table 3 shows the optimal hyperparameters obtained for all single-layer and multilayer networks used in the manuscript.

A2. Layer interdependence analysis

The layer interdependence problem consists of identifying which sets of layers are structurally related, and quantifying the strengths of those relationships. To this end, we use the MTCOV model and we employ the method described in De Bacco et al. (2017). This method consists of performing link prediction in one layer with and without the information in another layer to quantify the extent to which these two layers are related. Thus, for our purposes, interdependence is based on the idea that two layers are interdependent if the structure of one layer provides meaningful knowledge about the structure of the other.

To test our ability to predict a set of target layers α , we perform experiments with 5-fold cross-validation following the same routine as above by using only the optimal pair of hyperparameters. The main difference from the community-detection procedure above is the way the training and test sets are built. In fact, for the layer interdependence task, we only split (5-fold) the links in the set of target layers α together with the attributes for the nodes in this set, while giving full access to the set of layers β when they are added.

For this task, because we are mainly interested in link prediction, rather than in recovering covariates, we measure the AUC as in Equation (2). The final AUC is the average obtained over the five folds, each of which holds out a different subset of 20% of α . The value of the AUC depends both on the set of target layers α we are trying to predict, and on what set of other layers β we give the algorithm access to.

As described in Section 3.1, we restrict our analysis to the sublayers generated by the USPTO (separately) and the JPO and EPO layers (as sets of sublayers), without exploring all possible combinations of sublayers. In detail, we consider the following experiments:

- (a) $\alpha = [\text{US-APP}]$, $\beta_1 = [\text{US-EXM}]$, and $\beta_2 = [\text{US-EXM}, \text{EPO}, \text{JPO}]$.
- (b) $\alpha = [\text{US-EXM}]$, $\beta_1 = [\text{US-APP}]$, and $\beta_2 = [\text{US-APP}, \text{EPO}, \text{JPO}]$.
- (c) $\alpha = [\text{EPO}, \text{JPO}]$, $\beta_1 = [\text{US-APP}]$, $\beta_2 = [\text{US-EXM}]$, and $\beta_3 = [\text{US-APP}, \text{US-EXM}]$.
- (d) $\alpha = [\text{US-APP}, \text{EPO}, \text{JPO}]$, and $\beta_1 = [\text{US-EXM}]$.
- (e) $\alpha = [\text{US-EXM}, \text{EPO}, \text{JPO}]$, and $\beta_1 = [\text{US-APP}]$.

Table 3

Hyperparameters setting. Values of the hyperparameters C and γ extracted by 5-fold cross-validation combined with grid-search.

	US-EXM	US-APP	EP-APP	EP-ISR	EP-SEA	JP-REJ	JP-BCK	US-AGG	ALL-AGG	MULTI
C	7	7	7	7	3	7	7	7	7	15
γ	0.3	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7

Note that for the JPO and EPO, we are using all sublayers of these two jurisdictions. Furthermore, when the set α contains only a sublayer of USPTO [(a), (b)], the hyperparameters used by the algorithm are $C = 7$ and $\gamma = 0.7$, which is the optimal choice for the US-AGG network. For [(c), (d), (e)] the algorithm uses $C = 15$ and $\gamma = 0.7$, which is the optimal choice for the multilayer network, for computational simplicity.²⁸

Appendix B. Network comparison

B1. Class network construction

We use network comparison methods in order to quantify the differences in the technological information contained in the MULTI, ALL-AGG, and US-AGG networks. In particular, we directly compare a projection of the bipartite network of relationships between the extracted communities and the 3-digit Cooperative Patent Classification (CPC) classes with co-classification and inter-class citation networks. Fig. 5 displays the communities extracted for a random subset of the nodes and edges in ALL-AGG and US-AGG.

To build the bipartite network between communities and classes, we first populate a matrix P whose dimensions are given by the number of families (22653) times the number of classes (535). This is a binary matrix with non-zero entries when a family is assigned to a given class. We then normalize the matrix such that each column sums up to one. In this way, we can consider the matrix P to be the membership matrix of the classes among the patents. By multiplying the transpose of the membership matrix of the patents among the communities and the previous matrix P , we get the bipartite network $D = U^T P$ of relationships between the extracted communities and the classes. To ease the comparisons, we need to project this

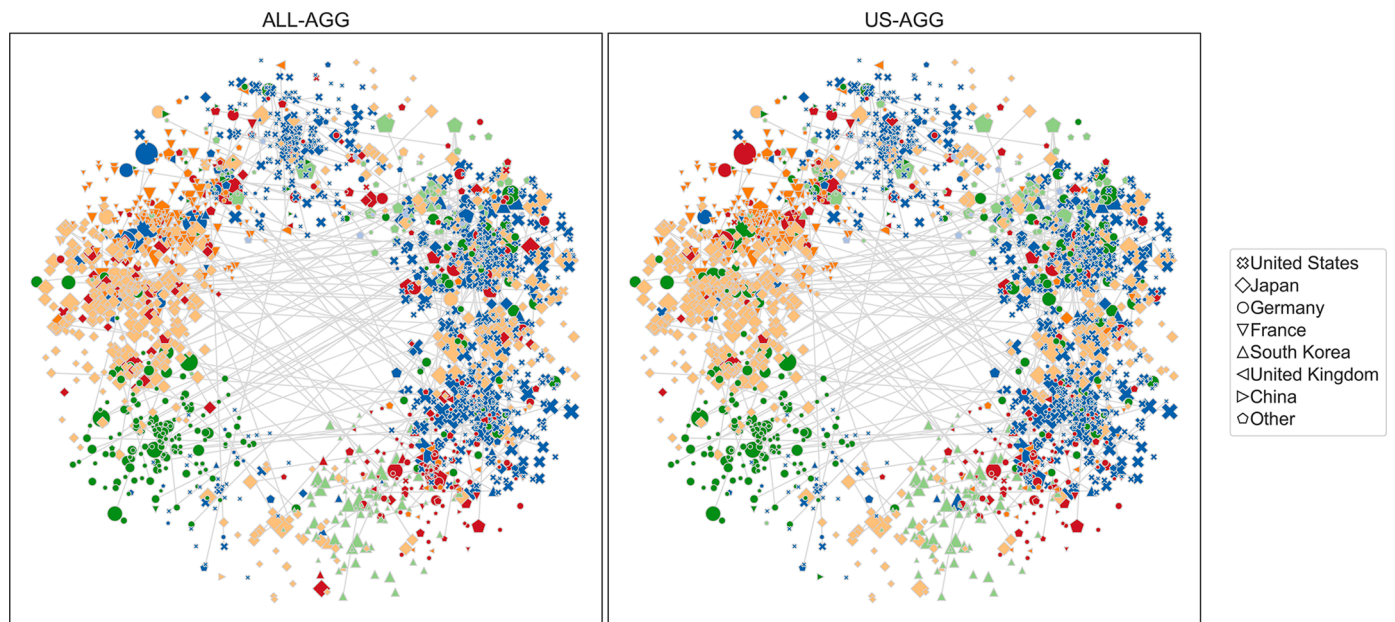


Fig. 5. Community extraction for comparison networks. This diagram shows the hard community membership partitions for the ALL-AGG and US-AGG networks. As for Fig. 4, we use a random sample of 2000 nodes and include any incidental links. The layout is determined by the results for MULTI, for purposes of direct comparison, while the colouring shows the communities found for each network (7 communities for each of ALL-AGG and US-AGG). Node size is proportional to the number of outgoing and incoming citations, while node shapes denote the location of the assignee of each patent family.

Table 4

Results of link prediction and covariate prediction tasks. We measure AUC (link prediction) and accuracy (covariate prediction) over 5-fold cross-validation for C equal to 7 (the optimal value for ALL-AGG and US-AGG) and 15 (the optimal value for the MULTI network); $\gamma = 0.7$ (the optimal value for all the networks).

	C	AUC	Accuracy
MULTI	7*	0.835	0.341
	15	0.852	0.422
ALL-AGG	7	0.730	0.402
	15*	0.739	0.393
US-AGG	7	0.736	0.426
	15*	0.749	0.406

²⁸ A cross-validation procedure to detect the best pair for the different sets α was determined to be too computationally expensive.

bipartite network onto the technology classes to obtain a network of classes. The projection onto the class nodes is computed through the matrix multiplication $D^T D$ between the bipartite matrix D and its transpose. This projection has non-zero entries when pairs of classes are both found in the same communities, with weights proportional to their relative frequencies within those communities. As baseline comparisons, we use the co-classification and the inter-class citation networks. The former is obtained by the matrix multiplication $P^T P$, while the latter is constructed as described in Section 3.3.

After running the community-detection algorithm for both the optimal and non-optimal partition of each of the three networks MULTI, ALL-AGG, and US-AGG, and only then can we obtain six projected networks among which we are able to make fair comparisons. Table 4 shows the performance of MTCOV on the citation networks (with non-optimal parameters identified with the symbol *) for the link prediction (AUC) and covariate prediction (accuracy) tasks, using 5-fold cross-validation.

After extracting communities, we construct the six bipartite (community-class) networks which we then project onto the class nodes to compare with the co-classification and inter-class citation networks.

B2. Diversity measures

Two diversity measures are used in the main body of this work: Rao-Stirling diversity (RSD) and the Herfindahl-Hirschman Index (HHI). For each network, RSD is calculated at the *extracted-community level* and then a median is taken across communities. The RSD for community c is calculated as (Stirling, 2007):

$$RSD_c = \sum_{i,j,i \neq j} d_{ij} p_{i,c} p_{j,c}, \quad (3)$$

where d_{ij} is a known distance measure between 3-digit CPC technology classes i and j , while p_i and p_j are the proportion of families in the community that are assigned classes i and j , respectively. Two factors complicate this calculation. First, because each family can be assigned multiple categories, RSD can take on values greater than one. Because we are directly comparing the RSD for the same set of families (our networks have the same set of nodes), this is not a concern. In fact, we believe this is sensible for this data. That is, if a community consists of a set of families that are all assigned the same two classes i and j , our procedure here will treat these communities as consisting of 100% i and 100% j (minimal diversity) rather than 50% i and 50% j (maximum diversity), for a given d_{ij} . Second, because we allow overlapping communities (i.e., a node can be assigned multiple communities with different weights), p_i and p_j are the weighted sums over patent families f in c :

$$p_{i,c} = \frac{\sum_{f \in i} w_{f,c}}{\sum_{\forall f} w_{f,c}}, \quad (4)$$

where $w_{f,c} \in [0, 1]$ is the weight of family f that is assigned to c .

For our purposes, d_{ij} is one minus the normalised link weight in the inter-class citation network constructed for our network comparison calculations. This metric is scaled such that distance zero corresponds to the strongest citation linkage for each class, and distance one corresponds to no citation linkage. These new weights act as proxies for the level of surprise, where a weight of zero indicates two classes that only ever cite each other, while a weight of unity indicates two classes that never cite each other. As such, the ‘level of surprise’ parameter d_{ij} down-weights combinations that we expect while exaggerating those that we don’t. This adjustment is important. For any given technology class, the number of classes with which it shares community membership depends crucially on both the classification system and the level of the hierarchy within this system that we choose to use. When a class starts to get too crowded, for example, it may be split to make technical search easier (Lafond and Kim, 2019) — after all, this is one of the primary goals of patent classification systems. For this reason, a distance measure like d_{ij} is crucial to incorporate into technological diversity measurements.

HHI, also called the Simpson diversity index, is calculated at the *technology level*, i , to measure the extent to which technology classes are split across extracted communities. A median across technology classes is then calculated. The HHI for class i is calculated as:

$$HHI_i = \frac{N \sum_c p_{i,c}^2 - 1}{N - 1}, \quad (5)$$

where $p_{i,c}$ is defined as in Equation (4) and N is the total number of communities into which families can be assigned (7 or 15, in our case). Equation (5) is the unbiased version of the HHI (Hall, 2005); this version corrects the $1/N$ offset that affects the standard version of the HHI (for which $1/N$ is the minimum value), which is the sum in the numerator of Equation (5). The HHI measures how much a technology class is splintered across communities, ranging from HHI=0 for maximally spread to HHI=1 for maximally concentrated. We note that the goal of the community detection process was *not* to replicate the CPC system as closely as possible. There are many valid reasons why a technology class may be split across communities, such as when a technology is particularly generalisable and is applied to (and cited by) many seemingly unrelated fields. Instead, the HHI gives us an idea of what is, or is not, driving the results we obtain for the direct network comparison.

Appendix C. Ego network details

The diagram in Fig. 1 shows the multilayer ego network of a triadic patent family, labelled A. Table 5 lists the seven families in this diagram, alongside their granted equivalents.

Table 5

Example network subset details. Details of the patent numbers within each of the families displayed in Fig. 1 are shown below. Priority indicates the month of first filing. All families consist of three triadic patents except for C, which includes multiple family members at the USPTO and EPO.




Node	DOCDB Family	Equivalent			Priority
		USPTO	EPO	JPO	
A	19192289	6819081	1333511	3671007	2002-01
B	17414436	6174618	0905803	3777748	1997-09
C	26411133	6211645, 6211646	0892450, 1030389, 1030390	4487967	1997-03
D	36242735	7615309	1695401	4527117	2003-12
E	37115311	7687192	1872418	4739405	2005-04
F	38522869	7967506	1994626	5133335	2005-03
G	37115315	7488201	1872421	4663781	2005-04

References

- Acemoglu, D., Akgicig, U., Kerr, W.R., 2016. Innovation network. *Proceedings of the National Academy of Sciences* 113 (41), 11483–11488.
- Aghion, P., Dechezleprêtre, A., Hémous, D., Martin, R., Van Reenen, J., 2016. Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. *Journal of Political Economy* 124 (1), 1–51.
- Alcácer, J., Gittelman, M., Sampat, B., 2009. Applicant and examiner citations in US patents: An overview and analysis. *Research Policy* 38 (2), 415–427.
- Aleta, A., Meloni, S., Moreno, Y., 2017. A multilayer perspective for the analysis of urban transportation systems. *Scientific Reports* 7 (1), 1–9.
- Almeida, P., Kogut, B., 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management Science* 45 (7), 905–917.
- Alstott, J., Triulzi, G., Yan, B., Luo, J., 2017. Mapping technology space by normalizing patent networks. *Scientometrics* 110 (1), 443–479.
- Asheim, B.T., Gertler, M.S., 2005. *The geography of innovation: Regional innovation systems.* The Oxford Handbook of Innovation. Oxford University Press.
- Azagra-Caro, J.M., Mattsson, P., Perruchas, F., 2011. Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. *Journal of the American Society for Information Science and Technology* 62 (9), 1727–1740.
- Bacchiocchi, E., Montobbio, F., 2010. International knowledge diffusion and home-bias effect: Do USPTO and EPO patent citations tell the same story? *Scandinavian Journal of Economics* 112 (3), 441–470.
- Bakker, J., Verhoeven, D., Zhang, L., Van Looy, B., 2016. Patent citation indicators: One size fits all? *Scientometrics* 106 (1), 187–211.
- Balland, P.-A., Rigby, D., 2017. The geography of complex knowledge. *Economic Geography* 93 (1), 1–23.
- Barbieri, N., 2016. Fuel prices and the invention crowding out effect: Releasing the automotive industry from its dependence on fossil fuel. *Technological Forecasting and Social Change* 111, 222–234.
- Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., et al., 2021. The physics of higher-order interactions in complex systems. *Nature Physics* 17 (10), 1093–1098.
- Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., Petri, G., 2020. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* 874, 1–92.
- Battiston, F., Nicosia, V., Latora, V., 2014. Structural measures for multiplex networks. *Physical Review E* 89 (3), 032804.
- Berkes, E., Gaetani, R., 2021. The geography of unconventional innovation. *The Economic Journal* 131 (636), 1466–1514.
- Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D., 2011. Foundations of multidimensional network analysis. 2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE, pp. 485–489.
- Bessen, J. E., 2003. Patent thickets: Strategic patenting of complex technologies. *Available at SSRN 327760*.
- Biddinger, B.P., 2000. Limiting the business method patent: A comparison and proposed alignment of European, Japanese and United States patent law. *Fordham L. Rev.* 69, 2523.
- Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., Zanin, M., 2014. The structure and dynamics of multilayer networks. *Physics Reports* 544 (1), 1–122.
- Breschi, S., Lissoni, F., Malerba, F., 2003. Knowledge-relatedness in firm technological diversification. *Research Policy* 32 (1), 69–87.
- Bródka, P., Kazienko, P., Musiał, K., Skibicki, K., 2012. Analysis of neighbourhoods in multi-layered dynamic social networks. *International Journal of Computational Intelligence Systems* 5 (3), 582–596.
- Cai, J., Li, N., 2019. Growth through inter-sectoral knowledge linkages. *The Review of Economic Studies* 86 (5), 1827–1866.
- Castaldi, C., Frenken, K., Los, B., 2015. Related variety, unrelated variety and technological breakthroughs: an analysis of US state-level patenting. *Regional Studies* 49 (5), 767–781.
- Choi, C., Park, Y., 2009. Monitoring the organic structure of technology based on the patent development paths. *Technological Forecasting and Social Change* 76 (6), 754–768.
- Chun, D., 2011. Patent law harmonization in the age of globalization: The necessity and strategy for a pragmatic outcome. *J. Pat. & Trademark Off. Soc'y* 93, 127.
- Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., Caldarelli, G., 2019. The statistical physics of real-world networks. *Nature Reviews Physics* 1 (1), 58–71.
- Clough, J.R., Gollings, J., Loach, T.V., Evans, T.S., 2015. Transitive reduction of citation networks. *Journal of Complex Networks* 3 (2), 189–203.
- Contisciani, M., Power, E.A., De Bacco, C., 2020. Community detection with node attributes in multilayer networks. *Scientific Reports* 10 (1), 1–16.
- Criscuolo, P., Verspagen, B., 2008. Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy* 37 (10), 1892–1908.
- Danguy, J., 2017. Globalization of innovation production: A patent-based industry analysis. *Science and Public Policy* 44 (1), 75–94.
- De Bacco, C., Power, E.A., Larremore, D.B., Moore, C., 2017. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E* 95 (4), 042317.
- De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M.A., Gómez, S., Arenas, A., 2013. Mathematical formulation of multilayer networks. *Physical Review X* 3 (4), 041022.
- De Domenico, M., Solé-Ribalta, A., Gómez, S., Arenas, A., 2014. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences* 111 (23), 8351–8356.
- De Domenico, M., Solé-Ribalta, A., Omodei, E., Gómez, S., Arenas, A., 2015. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications* 6, 6868.
- Dechezleprêtre, A., Fankhauser, S., Glachant, M., Stoeber, J., Touboul, S., 2020. *Invention and global diffusion of technologies for climate change adaptation.* Technical report. World Bank, Washington, DC.
- Demey, Y.T., Golzio, D., 2020. Search strategies at the European Patent Office. *World Patent Information* 63, 101989.
- Dernis, H., Khan, M., 2004. Triadic patent families methodology. Technical report. OECD.
- Engelsman, E.C., van Raan, A.F., 1994. A patent-based cartography of technology. *Research Policy* 23 (1), 1–26.
- Fink, C., Khan, M., Zhou, H., 2016. Exploring the worldwide patent surge. *Economics of Innovation and New Technology* 25 (2), 114–142.
- Fleming, L., 2001. Recombinant uncertainty in technological search. *Management Science* 47 (1), 117–132.
- Fleming, L., Sorenson, O., 2001. Technology as a complex adaptive system: Evidence from patent data. *Research Policy* 30 (7), 1019–1039.
- Fortunato, S., 2010. Community detection in graphs. *Physics Reports* 486 (3-5), 75–174.
- Frakes, M.D., Wasserman, M.F., 2017. Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. *Review of Economics and Statistics* 99 (3), 550–563.
- Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. *Management Science* 63 (3), 791–817.
- Gallotti, R., Porter, M.A., Barthelemy, M., 2016. Lost in transportation: Information measures and cognitive limits in multilayer navigation. *Science Advances* 2 (2), e1500445.
- Hall, B., 2005. A note on the bias in Herfindahl-type measures based on count data. *Revue d'économie industrielle* 110 (1), 149–156.
- Harvey, E., Maclaren, O., O'Neale, D., Patten-Elliott, F., Turnbull, S., Wu, D., 2021. *Network modelling of elimination strategy pillars: Prepare for it, stamp it out.* Technical report. Te Pūnaha Matatini.
- Haščič, I., Migotto, M., 2015. Measuring environmental innovation using patent data. OECD Environment Working Papers 89. OECD.
- Herfindahl, O.C., 1950. *Concentration in the steel industry.* Columbia University. Phd thesis.
- Higham, K., De Rassenfosse, G., Jaffe, A.B., 2021. Patent quality: Towards a systematic framework for analysis and measurement. *Research Policy* 50 (4), 104215.
- Higham, K., Yoshioka-Kobayashi, T., 2022. Patent citation generation at the triadic offices: Mechanisms and implications for analysis. *Available at SSRN 4022851*.
- Higham, K., Governale, M., Jaffe, A., Zülicke, U., 2017. Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. *Physical Review E* 95 (4), 042309.
- Higham, K., Governale, M., Jaffe, A., Zülicke, U., 2019. Ex-ante measure of patent quality reveals intrinsic fitness for citation-network growth. *Physical Review E* 99 (6), 060301.
- Hirschman, A.O., 1945. National power and the structure of foreign trade. Univ of California Press.

- Hirschman, A.O., 1964. The paternity of an index. *The American Economic Review* 54 (5), 761–762.
- Hötte, K., Jee, S.J., Srivastav, S., Knowledge for a warmer world: A patent analysis of climate change adaptation technologies. arXiv preprint arXiv:2108.03722.
- Huenteler, J., Schmidt, T.S., Ossenbrink, J., Hoffmann, V.H., 2016. Technology life-cycles in the energy sector—technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change* 104, 102–121.
- Hummon, N.P., Dereian, P., 1989. Connectivity in a citation network: The development of dna theory. *Social Networks* 11 (1), 39–63.
- Ibrahim, A.A., Lonardi, A., De Bacco, C., 2021. Optimal transport in multilayer networks for traffic flow optimization. *Algorithms* 14 (7), 189.
- Jaffe, A.B., de Rassenfosse, G., 2017. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology* 68 (6), 1360–1374.
- Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics* 108 (3), 577–598.
- Kauffman, S., Lobo, J., Macready, W.G., 2000. Optimal search on a technology landscape. *Journal of Economic Behavior & Organization* 43 (2), 141–166.
- Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2021. Measuring technological innovation over the long run. *American Economic Review: Insights* 3 (3), 303–320.
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A., 2014. Multilayer networks. *Journal of Complex Networks* 2 (3), 203–271.
- Koutra, D., Vogelstein, J.T., Faloutsos, C., 2013. Deltacon: A principled massive-graph similarity function. *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 162–170.
- Kuhn, J., Younge, K., Marco, A., 2020. Patent citations reexamined. *The RAND Journal of Economics* 51 (1), 109–132.
- Lafond, F., Kim, D., 2019. Long-run dynamics of the US patent classification system. *Journal of Evolutionary Economics* 29 (2), 631–664.
- Lambiotte, R., Rosvall, M., Scholtes, I., 2019. From networks to optimal higher-order models of complex systems. *Nature Physics* 15 (4), 313–320.
- Lee, S., Yoon, B., Lee, C., Park, J., 2009. Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change* 76 (6), 769–786.
- Lee, W.S., Han, E.J., Sohn, S.Y., 2015. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technological Forecasting and Social Change* 100, 317–329.
- Leten, B., Belderbos, R., Van Looy, B., 2007. Technological diversification, coherence, and performance of firms. *Journal of Product Innovation Management* 24 (6), 567–579.
- Li, R., Chambers, T., Ding, Y., Zhang, G., Meng, L., 2014. Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology* 65 (5), 1007–1017.
- MacGarvie, M., 2005. The determinants of international knowledge diffusion as measured by patent citations. *Economics Letters* 87 (1), 121–126.
- van der Marel, A., Prasher, S., Carminito, C., O'Connell, C.L., Phillips, A., Kluever, B.M., Hobson, E.A., 2021. A framework to evaluate whether to pool or separate behaviors in a multilayer network. *Current Zoology* 67 (1), 101–111.
- Mariani, M.S., Medo, M., Lafond, F., 2019. Early identification of important patents: Design and validation of citation network metrics. *Technological Forecasting and Social Change* 146, 644–654.
- Martinez, C., 2010. Insight into different types of patent families. *Science, technology and industry working paper*. OECD.
- Martínez, C., 2011. Patent families: When do different definitions really matter? *Scientometrics* 86 (1), 39–63.
- McCabe, S., Torres, L., LaRock, T., Haque, S.A., Yang, C.-H., Hartle, H., Klein, B., 2021. netrd: A library for network reconstruction and graph distances. *Journal of Open Source Software* 6 (62), 2990.
- Mejia, C., Kajikawa, Y., 2020. Emerging topics in energy storage based on a large-scale analysis of academic articles and patents. *Applied Energy* 263, 114625.
- Mewes, L., 2019. Scaling of atypical knowledge combinations in american metropolitan areas from 1836 to 2010. *Economic Geography* 95 (4), 341–361.
- Morris, R.G., Barthelemy, M., 2012. Transport on coupled spatial networks. *Physical Review Letters* 109 (12), 128703.
- Morrison, G., Giovanis, E., Pammolli, F., Riccaboni, M., 2014. Border sensitive centrality in global patent citation networks. *Journal of Complex Networks* 2 (4), 518–536.
- Nakamura, H., Suzuki, S., Kajikawa, Y., Osawa, M., 2015. The effect of patent family information in patent citation network analysis: A comparative case study in the drivetrain domain. *Scientometrics* 104 (2), 437–452.
- Nakamura, K., Sasaki, A., 2016. 先行技術文献情報開示要件の実証分析：特許審査への影響 [disclosure of information on prior art documents: Impacts on patent examination] (in Japanese). *Kokumin Keizai Zasshi [Journal of Economics & Business Administration]* 213 (1), 79–97.
- Nanumyan, V., Gote, C., Schweitzer, F., 2020. Multilayer network approach to modeling authorship influence on citation dynamics in physics journals. *Physical Review E* 102 (3), 032303.
- Newman, M.E., Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical Review E* 69 (2), 026113.
- Nicosia, V., Bianconi, G., Latora, V., Barthelemy, M., 2013. Growing multiplex networks. *Physical Review Letters* 111 (5), 058701.
- Okada, Y., Naito, Y., Nagaoka, S., 2018. Making the patent scope consistent with the invention: Evidence from Japan. *Journal of Economics & Management Strategy* 27 (3), 607–625.
- Oliphant, T.E., 2006. *A guide to NumPy*, 1. Trelgol Publishing USA.
- Omodei, E., De Domenico, M., Arenas, A., 2017. Evaluating the impact of interdisciplinary research: A multilayer network approach. *Network Science* 5 (2), 235–246.
- Parshani, R., Rozenblat, C., Ietri, D., Ducruet, C., Havlin, S., 2011. Inter-similarity between coupled networks. *EPL (Europhysics Letters)* 92 (6), 68002.
- Persoon, P.G., Bekkers, R.N., Alkemade, F., 2020. The science base of renewables. *Technological Forecasting and Social Change* 158, 120121.
- Petit, E., Van Pottelsberghe, B., Gimeno Fabra, L., 2021. Are patent offices substitutes? ECARES Working Papers.2021
- Porter, M.A., 2018. What is... a multilayer network. *Notices of the AMS* 65 (11).
- Rao, C.R., 1982. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 1–22.
- de Rassenfosse, G., van Pottelsberghe, B., 2009. A policy insight into the R&D–patent relationship. *Research Policy* 38 (5), 779–792.
- Shapiro, C., 2000. Navigating the patent thicket: Cross licenses, patent pools, and standard setting. *Innovation Policy and the Economy* 1, 119–150.
- Simpson, E.H., 1949. Measurement of diversity. *nature* 163 (4148).688–688
- Solá, L., Romance, M., Criado, R., Flores, J., del Amo, A.G., Boccaletti, S., 2013. Eigenvector centrality of nodes in multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23 (3), 033131.
- Solé-Ribalta, A., De Domenico, M., Gómez, S., Arenas, A., 2014. Centrality rankings in multiplex networks. *Proceedings of the 2014 ACM conference on Web science*, pp. 149–155.
- Sorenson, O., Rivkin, J.W., Fleming, L., 2006. Complexity, networks and knowledge flow. *Research Policy* 35 (7), 994–1017.
- Stirling, A., 2007. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface* 4 (15), 707–719.
- Sun, B., Kolesnikov, S., Goldstein, A., Chan, G., 2021. A dynamic approach for identifying technological breakthroughs with an application in solar photovoltaics. *Technological Forecasting and Social Change* 165, 120534.
- Tahmouresnejad, L., Beaudry, C., 2019. Capturing the economic value of triadic patents. *Scientometrics* 118 (1), 127–157.
- Tantardini, M., Ieva, F., Tajoli, L., Piccardi, C., 2019. Comparing methods for comparing networks. *Scientific Reports* 9 (1), 1–19.
- Taylor, D., Porter, M.A., Mucha, P.J., 2021. Tunable eigenvector-based centralities for multiplex and temporal networks. *Multiscale Modeling & Simulation* 19 (1), 113–147.
- Vaiana, M., Muldoon, S.F., 2020. Multilayer brain networks. *Journal of Nonlinear Science* 30 (5), 2147–2169.
- Valverde, S., Solé, R.V., Bedau, M.A., Packard, N., 2007. Topology and evolution of technology innovation networks. *Physical Review E* 76 (5), 056118.
- Vasques Filho, D., O'Neale, D.R., 2018. Degree distributions of bipartite networks and their projections. *Physical Review E* 98 (2), 022307.
- Veefkind, V., Hurtado-Albir, J., Angelucci, S., Karachalios, K., Thumm, N., 2012. A new EPO classification scheme for climate change mitigation technologies. *World Patent Information* 34 (2), 106–111.
- Verhoeven, D., Bakker, J., Veuglers, R., 2016. Measuring technological novelty with patent-based indicators. *Research Policy* 45 (3), 707–723.
- Verluise, C., Cristelli, G., Higham, K., de Rassenfosse, G., 2020. The missing 15 percent of patent citations. Available at SSRN 3754772.
- Verspagen, B., 2007. Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems* 10 (01), 93–115.
- Von Graevenitz, G., Wagner, S., Harhoff, D., 2011. How to measure patent thickets—a novel approach. *Economics Letters* 111 (1), 6–9.
- Von Wartburg, I., Teichert, T., Rost, K., 2005. Inventive progress measured by multi-stage patent citation analysis. *Research Policy* 34 (10), 1591–1607.
- Wada, T., 2016. Obstacles to prior art searching by the trilateral patent offices: Empirical evidence from international search reports. *Scientometrics* 107 (2), 701–722.
- Wada, T., 2020. When do the USPTO examiners cite as the EPO examiners? An analysis of examination spillovers through rejection citations at the international family-to-family level. *Scientometrics* 125 (2), 1591–1615.
- Wasserman, S., Faust, K., et al., 1994. *Social network analysis: Methods and applications*. Cambridge University Press.
- Wu, L., Wang, D., Evans, J.A., 2019. Large teams develop and small teams disrupt science and technology. *Nature* 566 (7744), 378–382.
- Yan, B., Luo, J., 2017. Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology* 68 (2), 423–437.
- Yuan, X., Li, X., 2020. A network analytic method for measuring patent thickets: A case of fcev technology. *Technological Forecasting and Social Change* 156, 120038.
- Yuvaraj, M., Dey, A.K., Lyubchich, V., Gel, Y.R., Poor, H.V., 2021. Topological clustering of multilayer networks. *Proceedings of the National Academy of Sciences* 118 (21).
- Zingg, C., Nanumyan, V., Schweitzer, F., 2020. Citations driven by social connections? a multi-layer representation of coauthorship networks. *Quantitative Science Studies* 1 (4), 1493–1509.
- Zingg, R., Fischer, M., 2018. The nanotechnology patent thicket revisited. *Journal of Nanoparticle Research* 20 (10), 1–6.

Generative model for reciprocity and community detection in networks

Hadiseh Safdari ^{*,†} Martina Contisciani ^{*,‡} and Caterina De Bacco [§]
 Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen 72076, Germany



(Received 14 January 2021; accepted 15 April 2021; published 14 June 2021)

We present a probabilistic generative model and efficient algorithm to model reciprocity in directed networks. Unlike other methods that address this problem such as exponential random graphs, it assigns latent variables as community memberships to nodes and a reciprocity parameter to the whole network rather than fitting order statistics. It formalizes the assumption that a directed interaction is more likely to occur if an individual has already observed an interaction towards her. It provides a natural framework for relaxing the common assumption in network generative models of conditional independence between edges, and it can be used to perform inference tasks such as predicting the existence of an edge given the observation of an edge in the reverse direction. Inference is performed using an efficient expectation-maximization algorithm that exploits the sparsity of the network, leading to an efficient and scalable implementation. We illustrate these findings by analyzing synthetic and real data, including social networks, academic citations, and the Erasmus student exchange program. Our method outperforms others in both predicting edges and generating networks that reflect the reciprocity values observed in real data, while at the same time inferring an underlying community structure. We provide an open-source implementation of the code online.

DOI: [10.1103/PhysRevResearch.3.023209](https://doi.org/10.1103/PhysRevResearch.3.023209)

I. INTRODUCTION

Reciprocity in directed networks, i.e., the tendency of a pair of nodes to form mutual connections between each other [1], is an important feature of many social relationships. Its impact ranges from affecting the development of exchange and power to determining the emergence of trust and solidarity [2,3]. Behavior of this kind has also been found in many kinds of networks that reflect human and institutional interaction, e.g., the world wide web, online dating, interfirm contracts, journal citations and email communication [4–8].

Among the various network modeling approaches, that of probabilistic generative models enable us for a rigorous theoretical foundation within the framework of statistical inference, as well as a flexible incorporation of domain knowledge in the modeling assumptions. Here, we consider a latent variable model, a probabilistic approach that contains latent and observed variables. The latent variables encode hidden patterns in the data, such as community memberships, and determine the probability of ties between nodes. For instance, knowing which communities two nodes belong to helps determine the likelihood of their interaction.

While in some simple cases, community structure may explain the tendency toward reciprocation [9], this mechanism may not be enough to capture more complex scenarios. Indeed, many generative models with community structure fail to reproduce the values of reciprocity observed in real networks, as we discuss in more details later. Conversely, several models aimed at capturing reciprocity do not account for community structure [10,11]. It is reasonable to expect that the mechanism regulating the existence of interactions can be influenced by both patterns of communities and reciprocity. In addition, communities are often interpretable objects and may correspond to functional unit, hence the value of including them in the model formulation. Incorporating reciprocity as well as community structure into a coherent latent variable model comes with the main challenge of relaxing the conditional independence assumption between edges, a common assumption in generative models to ease mathematical derivations. In addition, this task requires properly capturing conditional probabilities, as we describe later. Inspired by these insights, we propose a novel probabilistic latent variable approach to model networks that preserves the benefits of generative models, while capturing both community structure and reciprocity.

Models for reciprocity and latent community structure have largely been developed independently of one another, and only a handful of works have hinted at incorporating them into a unique framework. For instance, Garlaschelli and Loffredo [12] point towards a possible relationship between their model for reciprocity and general hidden variable models. Most notably, the pair-dependent stochastic block model of Holland *et al.* [9], well explained also by Wasserman and Anderson [13], holds assumptions similar to ours, in that it models jointly pairs of edges, which they call dyad vectors.

*These authors contributed equally to this work.

†hadiseh.safdari@tuebingen.mpg.de

‡martina.contisciani@tuebingen.mpg.de

§caterina.debacco@tuebingen.mpg.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by the Max Planck Society.

While a seminal work, it is, nevertheless, limited to hard membership and unweighted networks; hence the likelihood function that they propose substantially differs from the likelihood represented by our model. One practical aspect of our choice for the likelihood is that parameters' inference in our model is optimized to fully exploit the sparsity of the dataset and is scalable to large network sizes.

Reciprocity is often modeled by means of exponential random graphs [10,11,14,15], where it is treated as a measured network property that needs to be reproduced (often together with other network properties like the degree) by sampling networks using statistical mechanics principles, e.g., maximum entropy. The approach presented in this work significantly differs from the previous studies in that we include latent variables, such as community membership, as a mechanism to determine edge formation. However, in the case of exponential random graphs, possible group structures are not given a priori as the latent parameters; instead, they can only be estimated a posteriori on the sampled networks. More broadly, our approach is that of generative models, which incorporate a priori community structure by means of latent variables, and these are inferred from the observed interactions [16,17]. However, in these generative models, reciprocity is not explicitly included as a mechanism for tie formation, thus these models often fail to reproduce the observed reciprocity values of real networks. Consequently, a generative method whose latent variables describe both reciprocity and community memberships is needed.

II. RELAXING THE CONDITIONAL INDEPENDENCE ASSUMPTION

A possible explanation for the practical deficiency of generative models with communities to reproduce observed reciprocity values is the common assumption of conditional independence between edges, which makes the problem both analytically and computationally more tractable. This assumption states that the likelihood of a directed tie between two nodes depends only on their community membership (and other possible model parameters) but not on the existence of the reciprocated edge. This might be too strict of an assumption to capture the feature of reciprocity, where it is reasonable to expect that the existence of an edge in one direction should also be conditioned on the existence of an edge in the opposite direction. For instance, if an author i has cited another author j , this might predict the probability of j also citing i . At the same time, knowing the communities that the authors belong to, could also help estimating this probability. Mathematically, this can be translated to relaxing the assumption of conditional independence, which is the approach we take here.

Formally, we represent interactions between N individuals as a weighted asymmetric matrix A , with entries A_{ij} being the number (or weight) of interactions from i to j ; for instance, the number of favors or services that i does for j , or the number of times that i has endorsed j , e.g., as paper citations. Our model assigns a *joint* likelihood $P(A_{ij}, A_{ji}|\Theta)$ to edges involving the same pairs of nodes (i, j) , given some set of latent parameters Θ . Specifically, we assume the likelihood of a network to

factorize as

$$P(A|\Theta) = \prod_{i < j} P(A_{ij}, A_{ji}|\Theta) \quad . \quad (1)$$

This is fundamentally different from the prevalent approaches in generative models, where, typically, one assumes that *individual* edges are conditionally independent given the network parameters, i.e., $P(A|\Theta) = \prod_{i,j} P(A_{ij}|\Theta)$.

Notice that edges involving different pairs of nodes remain conditionally independent as in standard approaches. Equivalently, in terms of the conditional distribution of an individual edge $P(A_{ij}|A_{ji}, \Theta)$, we assume that this can be different than its marginal distribution $P(A_{ij}|\Theta)$. To the extent of our knowledge, this assumption has never been deeply questioned, except for a few works [18,19]. As firstly pointed out by Hoff [20], there are theoretical groundings for this assumption to hold in common scenarios, due to generalizations of de Finetti's theorem by Aldous [21] and Hoover [22] (see [19] for a detailed discussion). They show that, for exchangeable graphs, i.e., in networks without any natural order between nodes (which is often the case), the joint probability function of the adjacency entries can be properly represented using latent variables on nodes and pairs. In other words, the joint can be factorized as a product on edges, given the latent variables.

However, in the case of directed networks, where the adjacency matrix is asymmetric, as in our case, a precise representation can only be obtained using Eq. (1). While standard conditionally independent models can in principle arbitrarily well approximate the whole network distribution [23], in practice, it is not known how state-of-the-art models perform on this regard. To effectively model reciprocity, we relax the assumption of conditional independence and include the pairwise dependencies of two directed edges between pairs of nodes; such minimal relaxation is required to effectively model reciprocity. We compare results against standard conditionally independent models in terms of various performance metrics on both synthetic and real data.

III. THE COMMUNITY-RECIPROCITY MODEL

To fully specify the joint likelihood in Eq. (1), we need to characterize conditional distributions and one-point marginals like the distribution $P(A_{ij}|A_{ji}, \Theta)$ and $P(A_{ij}|\Theta)$. Here, we aim at capturing reciprocity, hence we assume that observed interactions exist because of two types of contributions: (i) the communities that nodes belong to, as in general community detection frameworks like the stochastic block model [9], and (ii) the fact that an individual that receives a directed interaction is more likely to reciprocate. In order to construct a model flexible enough to capture weighted networks and overlapping communities, we utilize a mixed-membership approach, similar to Refs. [16,17], to model how communities regulate edge formation.

Given the adjacency matrix A , our goal is to find community memberships of nodes and the magnitude of the reciprocity effect in the network, i.e., Θ . Bringing the contributions of reciprocity and community structure together, we model the conditional probability of A_{ij} given A_{ji} as drawn

from a Poisson distribution

$$P(A_{ij}|A_{ji}, \Theta) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{A_{ij}}}{A_{ij}!} \quad (2)$$

with mean

$$\lambda_{ij} = \lambda_{ij}^0 + \eta A_{ji} = \left(\sum_{k,q=1}^K u_{ik} v_{jq} w_{kq} \right) + \eta A_{ji}. \quad (3)$$

We denote with $\Theta = (u, v, w, \eta)$ the latent parameters that we want to infer. The parameters u_{ik}, v_{ik} are entries of K -dimensional vectors u_i and v_i , the out-going and in-coming communities, respectively; w_{kq} are the entries of a $K \times K$ affinity matrix, which regulates the structure of communities, e.g., assortative when its diagonal entries are greater than off-diagonal entries, in this case edges are more likely between nodes in the same community; η is the reciprocity parameter, and it regulates the impact of observing A_{ji} to predict the existence of A_{ij} . We omit from it the number of communities K , as in this work we assume this as given. When unknown, as in our experiments with real data, we estimate it by using cross-validation.

Notice that λ_{ij} includes separate contributions from both community parameters and reciprocity coefficient. It assumes additive contributions: we can have zero contribution from one term and still observe the existence of an edge because of the other term. If both are nonzero, their total impact sums up. This is conceptually different than a multiplicative contribution, a possible modeling choice that we do not explore here. Intuitively, an edge with weight A_{ij} exists if i and j belong to compatible communities (compatibility is regulated by the affinity matrix) or because of the reciprocity effect of observing the opposite edge A_{ji} . For instance, an author might cite another one because they belong to the same community (e.g., a research subfield) or because she was cited by the other on a previous publication.

Finally, as we need positive λ_{ij} , we assume $\eta \geq 0$. This restricts the model to have positive reciprocity contribution, i.e., receiving an in-coming edge can only boost the likelihood of the corresponding out-going edge, but not decrease it. Although this assumption could be limiting in certain contexts, it nevertheless applies to several relevant scenarios, in particular to the cases we study here. Relaxing this assumption, and suitably modifying the underlying theoretical model, is left for future works.

Our model specifies conditional probabilities, however, we do not assume the existence of a consistent joint distribution. In fact, finding a closed-form for the joint in Eq. (1), consistent with our proposed conditional, requires specifying a marginal probability function and then enforce consistency equations like $\sum_{A_{ji}} P(A_{ij}|A_{ji}, \Theta) P(A_{ji}|\Theta) = P(A_{ij}|\Theta)$. Depending on the choice of this marginal, enforcing consistency might be nontrivial, as it may require performing intractable marginalization. Early formalizations of the consistency between conditional and joint distribution has been provided, in

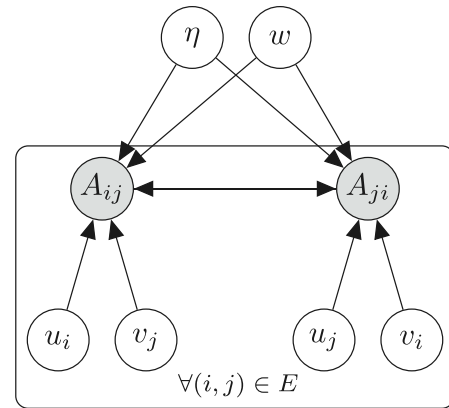


FIG. 1. Graphical model representation. A_{ij} and A_{ji} are the edges involving the same pairs of nodes (i, j) ; $\eta, w, u,$ and v are the latent parameters Θ ; E denotes the set of network edges.

a seminal work, by Besag's auto-Poisson models [24]. In the context of graphical models, a few models specify conditional Poisson distributions [25,26], but without considering latent variables. In the absence of a closed-form joint distribution, we adopt a tractable pseudolikelihood approach [24], where instead of optimizing the exact likelihood of Eq. (1), we consider the approximation

$$P(A|\Theta) = \prod_{i<j} P(A_{ij}, A_{ji}|\Theta) \approx \prod_{i,j} P(A_{ij}|A_{ji}, \Theta), \quad (4)$$

which is available in closed-form as it requires only the conditional probabilities, which we specified above. The equality holds only when A_{ij} and A_{ji} are conditionally independent, the common assumption in network generative models, as in that case $P(A_{ij}|A_{ji}, \Theta) = P(A_{ij}|\Theta)$. This is not our case since we relax this assumption, and Eq. (4) is only an approximation. This approach has also been considered in dyadic-dependent models [27], for community detection in networks [28], and for local Poisson graphical models [25]. A visual overview of our model is shown in Fig. 1.

IV. INFERENCE WITH EXPECTATION-MAXIMIZATION

The goal is to find the community and reciprocity parameters, i.e., Θ , given the adjacency matrix. Defining $L_{ij}^{ps}(\Theta, A_{ji}) = \ln P(A_{ij}|A_{ji}, \Theta)$ and neglecting the factorial term which is independent of these parameters, we have the log pseudolikelihood:

$$L^{ps}(\Theta) = \sum_{i,j} L_{ij}^{ps}(\Theta) = \sum_{i,j} (A_{ij} \ln \lambda_{ij} - \lambda_{ij}). \quad (5)$$

We aim at maximizing this quantity, but the presence of the logarithmic term makes this maximization difficult. However, using a variational approach by means of Jensen's inequality, it can be shown (see Appendix D 1) that maximizing $L^{ps}(\Theta)$

is equivalent to maximizing

$$L^{ps}(\Theta, \rho, \phi) = \sum_{i,j} \left\{ A_{ij} \rho_{ij}^{(1)} \left(\sum_{k,q} \phi_{ijkq} \ln u_{ik} v_{jq} w_{kq} - \sum_{k,q} \phi_{ijkq} \ln \phi_{ijkq} \right) + A_{ij} \rho_{ij}^{(2)} \ln \eta A_{ji} - A_{ij} (\rho_{ij}^{(1)} \ln \rho_{ij}^{(1)} + \rho_{ij}^{(2)} \ln \rho_{ij}^{(2)}) - \sum_{k,q} u_{ik} v_{jq} w_{kq} - \eta A_{ji} \right\}, \quad (6)$$

with respect to Θ , $\rho = (\rho^{(1)}, \rho^{(2)})$, and ϕ , where

$$\rho_{ij}^{(1)} = \frac{\lambda_{ij}^0}{\lambda_{ij}^0 + \eta A_{ji}}, \quad \rho_{ij}^{(2)} = \frac{\eta A_{ji}}{\lambda_{ij}^0 + \eta A_{ji}}, \quad (7)$$

$$\phi_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0}, \quad (8)$$

are the variational distributions over the parameters.

Constraints on the parameters can be arbitrarily added, e.g., $\sum_k u_{ik} = \sum_k v_{ik} = 1$, by incorporating Lagrange multipliers inside Eq. (5), and repeating similar calculations. In our numerical experiments, we consider both constrained and unconstrained cases.

We can perform this optimization by alternatively updating the various parameters, with an expectation-maximization (EM) algorithm. At each step, one updates ρ and ϕ using Eqs. (7) and (8) (E step) and then maximizes $L^{ps}(\Theta, \rho, \phi)$ with respect to Θ by setting partial derivatives to zero (M step). This iteration is repeated until L^{ps} converges. The

Algorithm 1 CRep: EM algorithm

Input: network $A = \{A_{ij}\}_{i,j=1}^N$,
number of communities K .

Output: membership vectors $u = [u_{ik}]$, $v = [v_{ik}]$; network-affinity matrix $w = [w_{kq}]$; reciprocity parameter η .

Initialize u, v, w, η at random.

Repeat until L^{ps} converges:

1. Calculate $\rho^{(1)}$ and ϕ (E step):

$$\rho_{ij}^{(1)} = \frac{\lambda_{ij}^0}{\lambda_{ij}^0 + \eta A_{ji}}, \quad \phi_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0}$$

2. Update parameters Θ (M step):

(i) for each node i and community k update memberships:

$$u_{ik} = \frac{1}{\gamma_i^u} \sum_{j,q} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}$$

$$v_{ik} = \frac{1}{\gamma_i^v} \sum_{j,q} A_{ji} \rho_{ji}^{(1)} \phi_{jikq}$$

(ii) for each pair (k, q) update affinity matrix:

$$w_{kq} = \frac{\sum_{i,j} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}}{\sum_{i,j} u_{ik} v_{jq}}$$

(iii) update reciprocity parameter:

$$\eta = \frac{\eta}{M} \sum_{i,j} \frac{A_{ij} A_{ji}}{\lambda_{ij}}$$

Note: γ_i^u, γ_i^v are quantities that are defined differently for constrained and unconstrained values of u_i and v_i . In the constrained case, they correspond to Lagrange multipliers; see Appendix D 2.

whole routine is described in Algorithm 1 and the detailed derivations are in Appendix D. This algorithm is computationally efficient and scalable to large system sizes as it exploits the sparsity of the dataset. Indeed, all the updates involve in the numerator sums over A_{ij} , hence only the nonzero entries count, giving an algorithmic complexity of $O(MK^2)$.

V. A BENCHMARK GENERATIVE MODEL WITH COMMUNITIES AND RECIPROCITY

So far we have focused on recovering the model parameters given the data, i.e., the inference. In this section, instead, we propose a benchmark probabilistic generative model to generate synthetic data with intrinsic community structure, and a given reciprocity value. It takes as input a set of membership vectors, u_i and v_i , affinity matrix w , and reciprocity parameter η ; the output is a directed network with adjacency matrix A . In this formulation, edges between a given pair of nodes are generated stochastically; one edge being generated first and independent from the other, while the formation of the opposite edge depends on how the first was drawn. The pairs of edges are conditionally independent from each other. Formally, we aim at sampling pairs of edges from Eq. (1), which can be done by selecting a marginal $P(A_{ij}|\Theta)$ and a conditional distribution $P(A_{ji}|A_{ij}, \Theta)$. By assuming a Poisson conditional as in Eq. (2) and a Poisson marginal, our model would reduce to a standard (conditionally independent) generative model with communities in the case of zero reciprocity parameter. Even though with this choice the joint is computationally intractable, this is not an issue, as we do not aim to use the joint to compute quantities analytically, but rather focus on sampling from it. Formally, given the input set of latent variables $\Theta = (u, v, w, \eta)$, we draw a pair (A_{ij}, A_{ji}) consistently with the joint $P(A_{ij}, A_{ji}|\Theta)$, in a two-step sampling routine:

(1) Select with a coin-flip one direction, (i, j) or (j, i) . Say we select (i, j) .

(2) Sample A_{ij} from the marginal

$$P(A_{ij}|\Theta) = \text{Pois}(m_{ij}), \quad (9)$$

where

$$m_{ij} = \frac{\lambda_{ij}^0 + \eta \lambda_{ji}^0}{(1 - \eta^2)} \quad (10)$$

is the mean of the marginal distribution such that it is consistent with the joint and the conditional distributions. Indeed, $\mathbb{E}[A_{ij}] = m_{ij} = \sum_{A_{ji}} A_{ij} P(A_{ij}|\Theta) = \sum_{A_{ij}, A_{ji}} A_{ij} P(A_{ij}, A_{ji}|\Theta)$ (see Appendix D 3 for the complete derivation).

(3) Sample A_{ji} from the conditional

$$P(A_{ji}|A_{ij}, \Theta) = \text{Pois}(\lambda_{ji}^0 + \eta A_{ij}), \quad (11)$$

using the previously extracted value of A_{ij} .

The Poisson distribution may generate multiple edges between a pair of nodes, so this model may create multigraphs. This is consistent with the interpretation that A_{ij} is the number, or total weight, of links from i to j . If we wish to generate binary networks where $A_{ij} \in \{0, 1\}$, we use the fact that the Poisson and Bernoulli distributions become close in the sparse limit. To enforce sparsity, it is sufficient to multiply the λ_{ij}^0 by a constant ζ , as the m_{ij} in Eq. (10) will also be automatically rescaled by the same quantity. The constant can be fixed by choosing a value for the expected number of (weighted) edges:

$$\mathbb{E}[M] = \sum_{i,j} \frac{\zeta \lambda_{ij}^0 + \zeta \eta \lambda_{ji}^0}{1 - \eta^2} = \frac{\zeta}{1 - \eta} \sum_{i,j} \lambda_{ij}^0 \quad (12)$$

$$\rightarrow \zeta = (1 - \eta) \frac{\mathbb{E}[M]}{\sum_{i,j} \lambda_{ij}^0}. \quad (13)$$

Imagine now a practitioner willing to control for the relative contribution of community and reciprocity in generating edges. Our model naturally allows this possibility, as this tuning is encoded by η . To see this explicitly, we calculate the fraction of edges generated by community effects only and introduce the cr_{ratio} variable as following:

$$cr_{\text{ratio}} := \frac{\sum_{i,j} \lambda_{ij}^0}{\mathbb{E}[M]} = 1 - \eta, \quad (14)$$

where we used Eq. (10) to rewrite the denominator. Thus, by varying η in the input, one automatically tunes the interplay community vs reciprocity: η close to 0 gives a network whose edges depend mostly on the community structure imposed by the membership vectors; instead, η close to 1 results in a network with lower impact of community structure, i.e., reciprocity has also significant impact on the edge formation. Notice that it is not possible to have a contribution purely due to reciprocity, as this phenomenon implicitly requires the existence of another mechanism to produce one of the two possible edges, here the community structure. This can also be seen by observing that Eq. (10) can be rewritten as $m_{ij} = \lambda_{ij}^0 + \frac{\eta}{1-\eta^2} (\eta \lambda_{ij}^0 + \lambda_{ji}^0)$; while the first term only depends on communities, the second term depends on both communities and reciprocity and they cannot be separated independently.

Having presented how our model can be used to generate synthetic data, we now proceed in describing how our model relates to observable network properties and how it can be used to predict reciprocated edges.

VI. PREDICTING NETWORK RECIPROCITY

In directed networks, reciprocity r is usually defined as the fraction of edges that are reciprocated [1], although other definitions exist to capture this feature [15,29]. With our probabilistic model, we can compute the expected value of a related quantity

$$r_w := \frac{\sum_{i,j} [A_{ij} A_{ji}]}{\sum_{i,j} [A_{ij}]}, \quad (15)$$

which corresponds to reciprocity in the case of binary adjacency matrices. A natural question is thus how this observable quantity is related to the reciprocity parameter η . In fact, we show that, provided some assumptions for the second moment $\mathbb{E}[A_{ij}^2]$ and considering an approximation with Taylor expansion (see Appendix D 4), η is a lower bound for it:

$$\mathbb{E}[r_w] \approx \eta + \frac{\sum_{i,j} [\lambda_{ij}^0 m_{ji} + \eta m_{ji}^2]}{\sum_{i,j} m_{ij}} \geq \eta. \quad (16)$$

The tightness of this bound depends on the latent variables through λ_{ij}^0 , (implicitly) m_{ij} , and m_{ji} . Empirically, we find that in the majority of the experiments the bound is very tight, i.e., $\mathbb{E}[r_w] \approx \eta$ and the other terms in Eq. (16) are much smaller than η in models with the conditional independence assumption, such as our proposed model with $\eta = 0$, where $\mathbb{E}[r_w] = \frac{\sum_{i,j} m_{ij} m_{ji}}{\sum_{i,j} m_{ij}}$. In fact, in these models, the term $\sum_{i,j} m_{ij} m_{ji}$ is often very small – we show empirical evidence of this later. Therefore, even in networks with high reciprocity, models with conditional independence assumption could poorly reproduce the term. This empirical result also seems to indicate that the pseudolikelihood approximation of Eq. (4) is relatively good in our datasets. The practical indication for practitioners is that networks generated by models with the conditional independence assumption have reciprocity values significantly different from those observed in real data.

VII. PREDICTING RECIPROCATED EDGES

The dependence structure between pairs of edges should allow us to predict the existence of a reciprocated tie *if* an edge in the opposite direction is observed, such as the citation of a paper if an author has been cited before by someone else. This is a kind of link prediction task, which lets us test the dependence assumption. It is also a principled way of comparing the accuracy of various generative models for any real network where no ground truth for the latent variables is known [30].

Conditional edge prediction can be formulated as follows: what is the probability of an edge $i \rightarrow j$ conditioned on observing the opposite existing edge (or nonexisting edge) $j \rightarrow i$? Our model naturally outputs this conditional probability. In contrast, a generative model that assumes conditional independence between edges is not capable of exploiting this extra information. It could only estimate marginal probabilities that do not depend on observing the opposite edge as it uses only the parameters such as community memberships and the affinity matrix. Our model is not capable of fully estimating marginal distributions but nevertheless can estimate its expected value as in Eq. (10). This is often the main quantity used in prediction tasks, as it plays the role of a score for estimating the entries A_{ij} . Therefore, with our model, we can also predict regular edge existence, where we simply aim at predicting an edge without any extra information but the inferred parameters.

In our experiments below, we test various generative models for both regular and conditional edge prediction by using 5-fold cross-validation. Specifically, we divide the dataset into five equal-size groups (folds) and give the models access to four groups (training data) for learning the parameters; this

contains 80% of the possible pairs of nodes in the network. One then predicts the existence of edges in the held-out group (test set). As performance metrics, we measure the AUC on the test data, i.e., the probability that a randomly selected edge has higher expected value than a randomly selected nonexisting edge. We compute both the regular AUC, by using as score the expected value $\mathbb{E}_{P(A_{ij}|\Theta)}[A_{ij}]$, and the *conditional* AUC (AUC–cond), which uses $\mathbb{E}_{P(A_{ij}|A_{ji},\Theta)}[A_{ij}]$ as the score, i.e., the expected value over the conditional distribution. The latter can only be computed for our algorithm, as for the others the marginal distribution is the same as the conditional, and thus the two AUC values coincide, see Appendix B for more details.

VIII. RESULTS

A. Results on real and synthetic data

We now demonstrate our model by applying it to both real and synthetic data. In the real-world datasets available to us, we only have a directed network of observed interactions, i.e., there is no available ground truth for the actual membership and reciprocity parameters. Consequently, their relative contributions in edge formation cannot be tuned. Thus we first validate our model and competing algorithms on synthetic data produced with different generative models. We test the ability of these models to (i) generate sample networks that replicate relevant network quantities such as reciprocity, similar to the observed values on the input networks; and (ii) perform edge prediction tasks. We then investigate our model’s performance on real-world datasets.

In the tests below, we use our model in various ways: the constrained version with constraints on the membership parameters u and v such that $\sum_k u_{ik} = \sum_k v_{ik} = 1, \forall i$ (CRep), the non constrained version (CRep_{nc}), and our model with $\eta = 0$ (CRep₀), i.e., without considering the reciprocity effect. For comparison, we use two generative models with latent variables: a community detection-only generative model with a maximum likelihood approach [16] (MT), which was the inspiration for the building block of our model in the case $\eta = 0$, and a Bayesian Poisson matrix factorization (BPMF) commonly used in recommendation systems [31]. For the edge prediction task on real data, we also consider a supervised learning link-prediction routine (OLP) with topological predictors and the implementation of Ghasemian *et al.* [32] (see Appendix G 3 for details).

B. Performance for synthetic networks

We study various types of synthetic networks, generated by three different models to cover several network topologies. Two of them cover the extreme scenarios of networks generated, accounting only for community structure or only for reciprocity. For the former, we use the standard stochastic block model (SBM) [9] and for the latter the reciprocity model of Holland and Leinhardt (HL) [10]. Our model, instead, is designed to tune the relative impact of community structure and reciprocity in determining edges, by varying the parameter η . Thus we use the benchmark generative model described above to interpolate between these two extremes by tuning η : for small values we reproduce the results equivalent to the

stochastic block model, whereas for higher values we replicate a structure similar to Holland and Leinhardt’s model.

The generative process is described in detail in Appendix A. As a remark, the exact joint likelihood of CRep is not determined in closed-form, however all the models used here for comparison adopt either its Poisson conditional distribution (our model with $\eta > 0$) or its Poisson marginal distribution (all the other models). Thus experiments here are aimed at highlighting differences in the various models’ assumptions. By varying the network sparsity and the impact of communities and reciprocity, we illustrate types of structure that may exist in real-world data, and test each algorithm’s robustness against them on various tasks including edge prediction and the ability to reproduce sample networks that replicate relevant network quantities.

Reproducing the topological properties. An important property of a model is the ability to generate network samples that resemble what is observed in real data. We test this ability by considering topological properties like degree distribution, reciprocity, and hierarchical structure. We calculate their values on network samples, which are generated with the various generative models, by applying the inferred parameters from the given input data. Specifically, we consider networks generated synthetically as explained above, and for each individual network we infer the parameters by each model, and use them to generate five network samples. We compare topological properties of these samples with those observed on the ground truth networks used to infer the parameters.

In particular, we are interested in measuring reciprocity, as the networks generated by algorithms only based on community structure are not capable of reflecting the observed value of the reciprocity in the ground truth network, a shortcoming of these models which indeed limits their applications. The empirical evidence of this observation was part of the motivation to study this problem. In the experiments, we use the standard definition of reciprocity r , i.e., the ratio of the number of edges pointing in both directions to the total number of edges in the graph (we use the PYTHON implementation in NETWORKX). As anticipated, in networks generated with the stochastic block model, r is often close to 0. Instead, a more interesting scenario is that of networks generated with the main purpose of replicating reciprocity, as in the HL model. This is an example of an exponential random graph model where reciprocity and sparsity are the two topological properties controlled in input. It is also one of the few cases where this type of model is analytical, see Appendix E. In this model, r is tuned by a parameter α so that the higher its value, the higher the reciprocity. Notice that, as usual in exponential random graphs models, latent variables such as communities are not considered. This model generates unweighted networks, hence $r \equiv r_w$.

Figure 2 shows that CRep significantly outperforms all the other generative models in reproducing r_w , panel (a), and r , panel (b), as measured on the sampled networks. The gap between the values of r and r_w on the sampled networks is due to the mismatch between the binary adjacency matrices of the networks generated with the HL model (input data) and the weighted sampled ones generated with the various generative models, which use Poisson distributions. Similar results are obtained for the networks generated with our benchmark

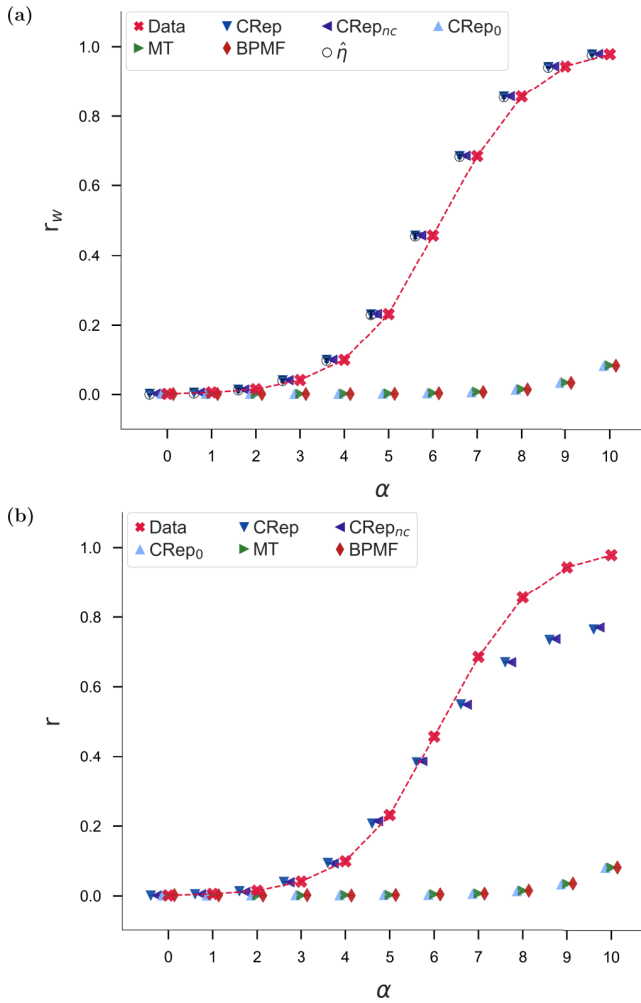


FIG. 2. Reciprocity in HL networks. Synthetic networks with $N = 1000$ nodes generated with the model proposed by Holland and Leinhardt by varying the reciprocity parameter α . Results are empirical averages and standard deviations over 15 samples of three independent synthetic networks (five samples per input network). The red markers indicate the average on the three input networks. (a) The quantity r_w as defined in Eq. (15); the empirical average over the samples and the theoretical expectation as in Eq. (16) coincide, hence we omit the markers for the empirical value; $\hat{\eta}$ is the inferred parameter in CRep and CRep_{nc}. (b) Standard reciprocity r . Notice that $r \equiv r_w$ for the input data, but this is not true for the samples, as the generative models considered here generate weighted edges, i.e. the matrix A is in general not binary. Error bars are smaller than marker size. Unless otherwise stated, this will be the case in all of the figures.

generative model. Also in this case, CRep captures reciprocity significantly better than the other models, consistently over a range of values of η as the input parameter. Moreover, in the case of fixed η , varying the sparsity and degree of overlapping communities lead to the same results. We leave details in Appendix F 1.

At this point, we turn our attention to topological properties other than reciprocity, to investigate how these generative models perform in reproducing various relevant properties that might be of interest for a practitioner. Indeed, other pos-

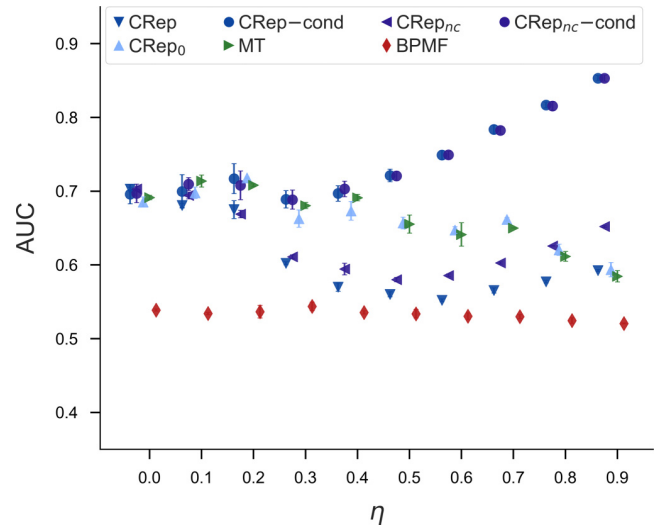


FIG. 3. Edge prediction in benchmark networks. Synthetic networks with $N = 2100$ nodes and $K = 3$ communities of equal-size unmixed group membership generated with the benchmark generative model proposed above by varying the reciprocity parameter η . The results are averages and standard deviations over three independent synthetic networks and over 5-fold of cross-validation test sets. The accuracy of edge prediction is measured with AUC and the baseline is the random value 0.5.

sible mechanisms underlying network interactions are those that involve more than two individuals (which is the case for reciprocity), e.g., hierarchical structure, which requires the whole network for its computation.

As in our experiments we find that all models are able to retrieve the degree distribution with good accuracy, we mainly focus on replicating ranking of nodes, an application relevant when nodes have a score representing some intrinsic notion of relative strength or prestige. For this, we use SPRINGRANK [33], an algorithm for inferring hierarchies in directed networks that assigns real-valued scores to nodes. We calculate the Gini index on these scores to provide a global measure for the whole network. Comparing the average over the five samples, we find that CRep and CRep₀ are able to perfectly retrieve the Gini index of the original network, while the other models tend to overestimate it, see Appendix F 1. This is consistent over the various synthetic network topologies. Notice that this topological property is influenced neither by the value of η , nor the fraction of nodes with mixed-membership used to generate networks; however, it decreases as the average degree, and α increase.

Edge prediction. We test the algorithms' ability in edge prediction tasks, in both cases of conditional and regular edge prediction. As we can see from Fig. 3, our model outperforms the others in conditional edge prediction, showing that it is able to efficiently exploit the additional information about the existence of the opposite edge. The performance gap between different approaches increases with η , as for high values of η , the reciprocity plays a bigger role in edge formation. In the opposite scenario of low η , the impact of reciprocity becomes negligible compared to community structure, and in this case we reproduce the same results as for the other algorithms. This

is expected as our model infers small values of η in this case, thus in practice reducing to a conditional independent model as the others. Performance in terms of regular edge prediction is comparable to the other algorithms for small η , while it drops for intermediate values and then increases again as η grows.

These synthetic tests suggest that working with conditional probabilities results in more robust estimates of the probability that an edge exists if we have access to the edge in the opposite direction. Performance improvement is more significant when community structure is not the predominant mechanism in edge formation. We leave more details in Appendix F 2.

To summarize results on synthetic networks, CRep is capable of suitably capturing the reciprocity values observed in a given network, while also retrieving hierarchical structures. Furthermore, CRep exploits the availability of extra information in performing edge prediction, by increased performance and robustness across various parameters' ranges.

C. Performance for real networks

Above, we evaluated the ability of our model, CRep, to generate network samples that have reciprocity values as expected in input and tested its performance in edge prediction. In this section, we examine these abilities on real world datasets. We apply our method to datasets from a diverse set of fields, with sizes ranging up to $N \sim 10^4$ nodes and up to $E \sim 10^5$ links (see Table I and Appendix G 1 for details). Together, these examples cover various types of social relationships, communication interactions, transportation systems, and patterns of citations.

Reproducing the topological properties. We apply the same procedure as before to infer the parameters $\Theta = (u, v, w, \eta)$ from data (this time, real networks) and then generate synthetic network samples based on them. Also in this case, CRep greatly outperforms the other models in reproducing r , consistently across datasets. We show as an example in Fig. 4 the results on the Erasmus dataset (Erasmus Mobility Network 2014–2018) [34], and we leave the others in Appendix G 2.

Previously, we have discussed network-related quantities controlled by η , such as the expected fraction of edges purely due to communities (cr_{ratio}) or the quantity r_w . Here we illustrate how the various real networks differ in the inferred values of η , which we denote as $\hat{\eta}$. In particular, we show in Fig. 5 how $\hat{\eta}$ varies according to the reciprocity of these networks, unveiling a nontrivial pattern. While we see a general trend of $\hat{\eta}$ increasing with r , there are interval ranges of r for which $\hat{\eta}$ varies widely across networks, and vice-versa. For example, we see that for $r \in [0.6, 0.8]$, $\hat{\eta}$ ranges in $[0.1, 0.7]$. This high variability suggests that r is the result of a complex combination of communities and reciprocity. We notice, for instance, that for high school friendship networks (HST and DT), $\hat{\eta}$ is low (i.e., in $[0.1, 0.3]$), showing that many reciprocated edges are explained by community structure. Instead, for online dating (POK) and communication networks (EU and DNC), we observe high values of $\hat{\eta}$, signaling a lower impact of communities, as reciprocity plays a bigger role. This reinforces the need to include in network models both mechanisms for explaining edge formation. Notice that these

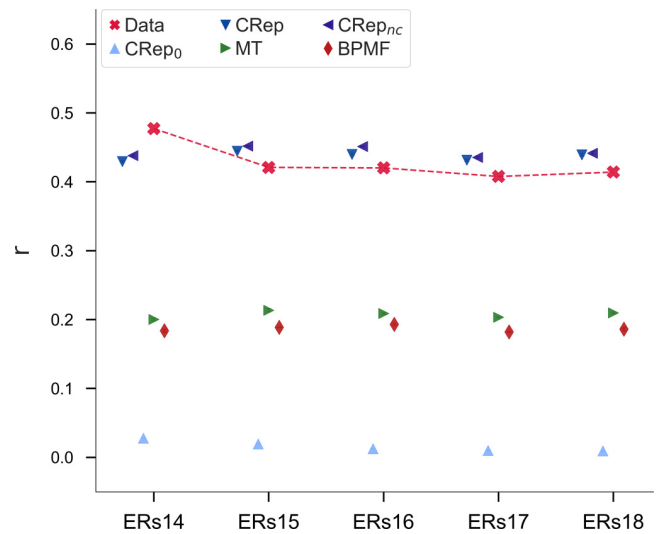


FIG. 4. Reciprocity in the Erasmus datasets. Results are averages and standard deviations of r over five samples generated with the various generative models. The algorithms use the inferred η and community parameters of the dataset—Erasmus in this plot—to generate synthetic network samples. Red markers indicate the values of r in the real datasets.

results are possible not only because our model accounts for reciprocity through an explicit parameter η , but also because it infers reciprocity values close to the observed ones, while the other methods fail at this, see Fig. 12.

Edge prediction. In the absence of ground truth, as in most real world networks, we test the ability in edge prediction by cross-validation, as done for synthetic networks. Table II shows the results in terms of AUC for the generative models CRep, MT, BPFM, as well as for OLP; the latter is a type

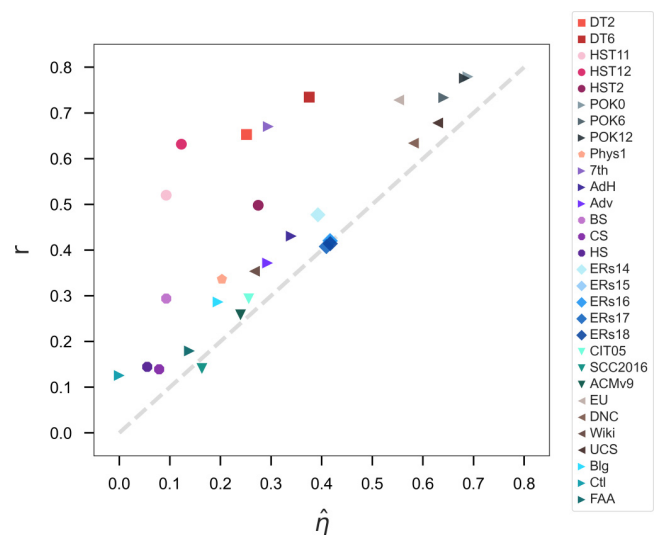


FIG. 5. Reciprocity and $\hat{\eta}$. Scatter plot with observed reciprocity (y axis) and $\hat{\eta}$ inferred in CRep (x axis); points are individual real datasets. The dashed grey line indicates the perfect correspondence between r and $\hat{\eta}$. Marker shape denotes the type of network as defined in Table I.

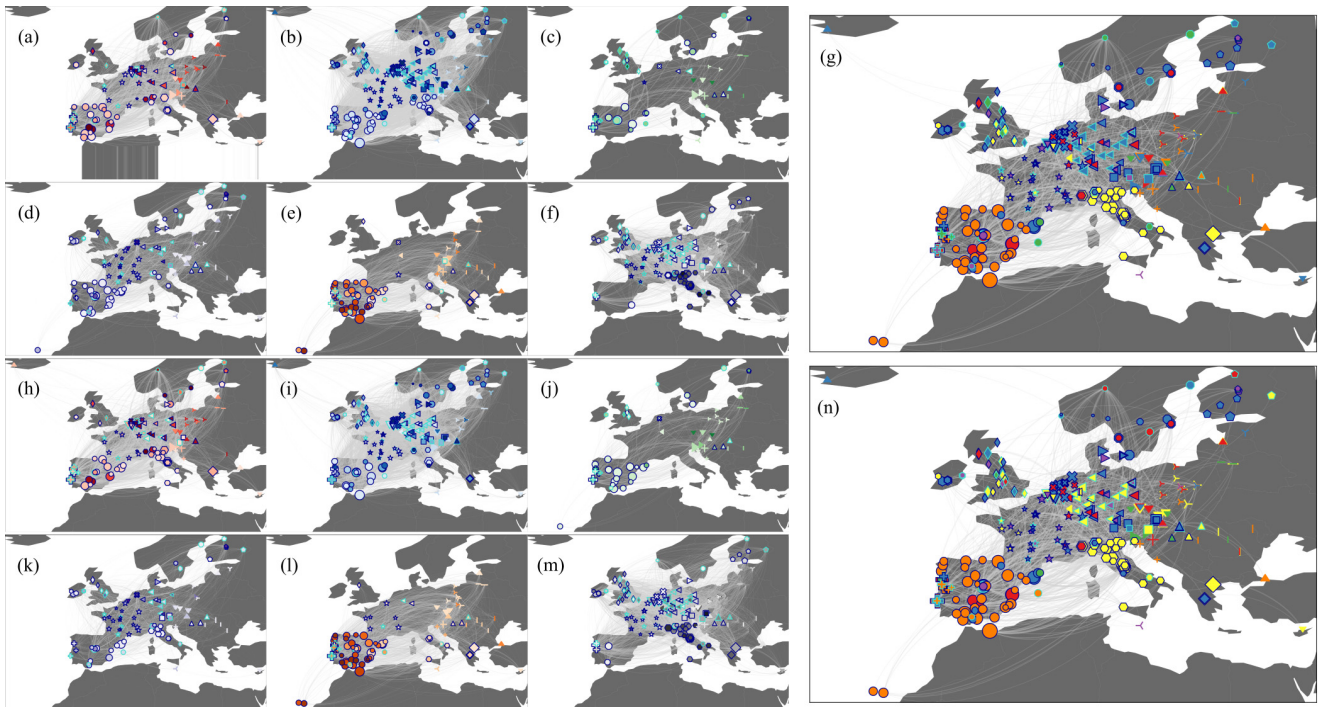


FIG. 6. Erasmus 2018 community structure. For visualization clarity, we show the subnetwork made of the 10% biggest institutions and the 3000 edges with highest weights (inference was performed on the whole network). (a)–(f) show groups $K = 1, 2, \dots, 6$ (mixed membership) by CRep_{nc} , and (h)–(m) show the same groups by MT. (g) illustrates the groups by CRep_{nc} in the case of hard membership, while the groups by MT are represented in (n). Node color intensity increases with u_{ik} , so that darker nodes have stronger membership u in that group, each color is a group (mixed membership) and nodes with light blue border are nodes that change the most the membership in the two algorithms; for each group k , we only show nodes that have $u_{ik} > 0.1$. Node and edge size are proportional to the size of an institution measured by the total number of outgoing and incoming students. Node shapes denote country.

of supervised learning technique which uses network topological information as features to predict the entries of A . CRep and OLP show the best results, with CRep having high performance for social networks. However, if we consider the conditional AUC, then CRep outperforms all the others in the majority of the datasets, as also observed in synthetic data. Finally, by averaging the AUC across the dataset, we find CRep_{nc} is the best model. This confirms the ability of our model to efficiently exploit the additional information from the adjacency matrix to boost performance in terms of edge prediction.

IX. CASE STUDY: APPLICATION OF CREP TO THE ERASMUS STUDENT EXCHANGE NETWORK

We illustrate our model on a real dataset to show various analysis that a practitioner can perform. We consider a network representation of the Erasmus student exchange program in 2018 [34], denoted as ERs18 in Table I. A node represents a higher education institution and an edge between nodes i and j denotes how many students were sent from i to spend a portion of their academic year abroad at institution j , as part of their study program towards a degree (Bachelor, Master, or PhD). This program is supported by the European Commission and involves $N = 4389$ institutions (mainly European), with a total of $M = 90\,972$ participating students in 2018.

We recover community partitions from the network data using both CRep_{nc} and MT, they have similar and high performance in edge prediction according to AUC (see Table II), and we fix $K = 6$ communities from cross-validation. In Fig. 6, we notice that while both models find several groups that closely correlate with countries, CRep_{nc} tends to put German institutions (left triangles) more in the same group (blue) and shifts few institutions in the red group, which seems made of mainly universities with strengths in engineering and technology (e.g., Universitat Politècnica de Catalunya, Politecnico di Milano, and Institut Polytechnique de Grenoble). For instance, Università di Bologna, Federico II di Napoli and Padova have lower $u_{i,red}$ than what is predicted without accounting for reciprocity, instead Slovenská technická univerzita v Bratislave, Kauno Technologijos Universitetas, and Universidad de Oviedo increase their membership in this group.

In addition, CRep_{nc} places more institutions with higher membership in the green group, see Fig. 6(g) (hard membership). While there is no apparent common attribute between these (e.g., country), we find that many nodes with high “green” entry of u_i tend to reciprocate more edges. Specifically, they have a high fraction of out-neighbors such that λ_{ij}^0 is much smaller than λ_{ji}^0 . That is, the edges A_{ij} such that A_{ji} also exists, have a lower impact in determining the value of u_i in the algorithm. In fact $u_{ik} \propto \sum_{j,q} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq} =$

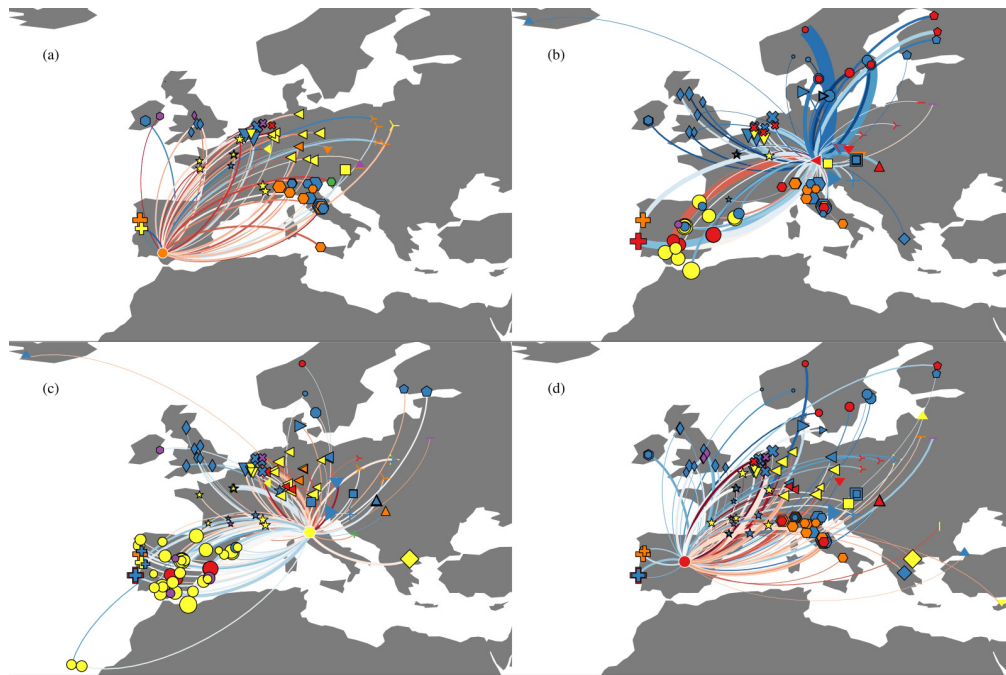


FIG. 7. Patterns of reciprocated edges. Plots show the subnetwork made of the reciprocated edges of (a) Universidad Pablo de Olavide, (b) Technische Universität München, (c) Università degli Studi di Firenze, and (d) Universidad Carlos III de Madrid. Node size is proportional to university size, the shape denotes country, the colors are the highest entry of u_i (for the four reference nodes - white node border) and v_j (for all its neighbors). Edge size is proportional to its weight; edge colors vary continuously from red to blue, based on the value of $d_{ij} = cr_{ij} - cr_{ji}$: high intensity red, white, and high intensity blue mean close to -1 , 0 , and 1 , respectively.

$\sum_{j,q} \frac{A_{ij} u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0 + \eta A_{ji}}$, see Eq. (7). Hence, if the denominator is high because of A_{ji} , the weight of the edge A_{ij} decreases. Nodes with many such A_{ij} tend to have lower entries u_{ik} and thus lower λ_{ij}^0 . This is a qualitative explanation for having different membership, however the situation is more complicated than this, as one needs to account for the effects on the whole network. In fact, also v_{jq} changes between the two algorithms, for a similar reason, thus also contributing to a different u_{ik} .

The primary benefit of CRep, however, lies not in its ability to recover the communities but in what it reveals about the reciprocity patterns in the network. Home and receiving institutions must sign an inter-institutional agreement to allow for student exchanges between them. While institutions may sign them because of clear affinities between their educational training offerings (e.g., both universities are strong in natural science), they might also do so because of some mechanisms involving reciprocity, as hosting students costs resources. Moreover, reciprocity could be further increased by previous knowledge or collaborations between individual faculties, thus institutional reciprocity may be also driven by faculty reciprocity. In addition to the communities themselves, our model also returns η , which can reveal features of the data related to such reciprocity effects not seen with standard generative models, such as cr_{ratio} or $\mathbb{E}[A_{ij}|A_{ji}, \Theta]$. We find a maximum likelihood value of $\eta = 0.4$, signaling a significant reciprocity effect. In fact, according to Eq. (14), on average 40% of the edges are influenced by reciprocity.

While η gives a global picture of the whole network, our models still allows to distinguish the impact of reciprocity

on individual edges. For instance, if an institution i accepts many students from j , then j might be more willing to accept students from i , even though i 's features might not match j 's preferences. If we distinguish the u_i as the set of preferences of i and v_j as the set of attributes of j , then our model will naturally convey this through high λ_{ij}^0 and low λ_{ji}^0 for such a case. CRep is able to capture these situations quantitatively, by means of the quantities $cr_{ij} := \lambda_{ij}^0/m_{ij}$ (a cr_{ratio} per edge) with values in $[0,1]$ which measures the relative contribution of communities alone to determine edges between i and j . Focusing on a single institution i , one can analyze the difference $d_{ij} := cr_{ij} - cr_{ji} \in [-1, 1]$ for all j such that both $A_{ij}, A_{ji} > 0$ and find different reciprocity patterns, as we show in Fig. 7. Here we plot three extreme cases where i has most of the d_{ij} being less, equal or greater than 0. The Universidad Pablo de Olavide in Sevilla, panel (a), has mostly $d_{ij} < 0$ (plotted in red), meaning that reciprocity has a strong effect in determining its out-going edges to universities that instead send students to Sevilla mostly out of community preference. The opposite case is that of Technische Universität München, panel (b), which has most of the $d_{ij} > 0$ (plotted in blue), signaling that it tends to select its out-going edges more out of preference than their counterparts, who tend to reciprocate instead. Università degli Studi di Firenze, panel (c), is an example of an institution with several d_{ij} close to 0 (plotted in white), meaning that most of its reciprocated edges are due to community affinities. In other words, Firenze selects out-going j based on preference and those who select Firenze do the same, so the impact of reciprocity is low. Apart from these three extremes, many universities display a range of

such behaviors; we give an example of Universidad Carlos III de Madrid, panel (d), which has a balanced fraction of reciprocated edges covering these three cases (there are about 1/3 of blue, red, and white edges in the corresponding figure). Notice that the value of d_{ij} yields an incomplete picture of the situation, since it does not distinguish between cases where the quantities cr_{ij} , cr_{ji} have different magnitudes while keeping their difference constant.

X. CONCLUSION

CRep is a mathematically principled generative model for capturing both community and reciprocity patterns in directed networks. It relies on relaxing strict conditional independence assumptions on edges that limit the applicability of standard methods on real problems where reciprocity plays an important role. Its algorithmic implementation is efficient and scalable to large system sizes. The corresponding generative model allows for the creation of synthetic networks with the desired interplay between community and reciprocity in determining the edges, while allowing the tuning of network sparsity.

In addition to providing all the analysis tools typical of standard generative models with communities, our model makes it possible to answer questions about reciprocity in networks that were not previously possible; for instance, performing probabilistic conditional edge prediction and estimating the relative contribution of community and reciprocity in determining edges. We show how real networks display a wide range of the reciprocity parameter, signaling the variety of possible patterns for this property. In the context of the Erasmus student exchange network, our model allowed us to distinguish universities based on their pattern of reciprocated edges.

More generally, our model shows how we can relax strict conditional independence assumptions on edges and showcases possible consequences in doing this. This presents an opportunity for researchers to rethink the fundamental assumptions behind generative models, and present models that may open doors to new theories and questions. We make one step in this direction, as our model connects two popular problems that are mainly treated independently: the inference of communities in networks and generating directed networks where reciprocity plays a relevant role. We used this connection to obtain networks with community structure and values of reciprocity consistent with those observed in real data.

Both the assumption and the model we have presented are only the first step in a broader line of work that investigates how certain topological properties are reflected in networks with latent community structure as dominant mechanism in edge formation. There are a number of directions in which this work could be extended. We have considered here a simple way to account for reciprocity and break conditional independence, by considering a unique parameter for the whole network. Our model could be extended to account for node-dependent parameters, where reciprocity varies between individuals. In addition, possible extensions may incorporate extra information such as degree, attribute or signals on nodes [30,35–38], edges of different types as in multilayer networks [16] and dynamics in time [39–44]. Reciprocity is one of

the many effects that could play a role in determining how nodes interact in a network. One could go further than this by considering incorporating quantities that account for triples of individuals, for instance clustering coefficient, transitivity or global centrality measures [45]. These properties cannot be captured by standard SBM-like models [46]. In this respect, a recent work of Peixoto [47] shares some similarities with ours considering triadic closure instead of reciprocity, making an effort towards extending the stochastic block model framework to incorporate more elaborate topological structure that is not captured otherwise. This is something that exponential random graphs or stochastic actor oriented models are capable of [14,48–51], without including latent community structure but rather fitting network statistics. In probabilistic generative models, this would require further breaking conditional dependencies between edges, potentially increasing the model complexity to encompass more complicated situations. With our work, we made the first step in this direction.

While there is no unique generative model that captures all the possible network properties well, our work illustrates how to target reciprocity. As our original motivation to study this problem came from the realization that standard generative models fail to generate synthetic networks with meaningful values for this property, our work illustrates a way in which latent variable frameworks can be applied more realistically, and provides an example of how network scientists can better align fundamental theories with realistic applications. We provide an open source implementation of the code online in Ref. [52].

ACKNOWLEDGMENTS

The authors thank Eleanor Power and Elspeth Ready for useful conversations. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani. All the authors were supported by the Cyber Valley Research Fund.

APPENDIX A: SYNTHETIC NETWORK GENERATION: NUMERICAL IMPLEMENTATION

The synthetic networks used in the analysis are of three types and represent different scenarios: networks with community structure only, with reciprocity only and networks with both communities and reciprocity. In order to obtain networks with only a community structure we use a stochastic block model with different values of average degree $\langle k \rangle$. We generate networks with $K = 3$ communities of equal-size un-mixed group membership, $N = 2100$ nodes and an assortative structure (w has higher diagonal entries) with main probabilities $p_1 = cK/N$ and entries outside the main diagonal equal to $p_2 = 0.1p_1$, so that the average degree is $\langle k \rangle = c + (K - 1)c/10$, where c is the average degree within the same community. We generate three independent samples for each value of $c \in [2, 20]$, that corresponds to $\langle k \rangle \in [2.4, 24]$. On the other hand, we generate networks influenced by reciprocity only through an implementation of the reciprocity model proposed by Holland and Leinhardt (see Appendix E for details). The input parameter α can be tuned to obtain different values of network reciprocity and we generate three

independent samples for each value of $\alpha \in [0, 10]$. We consider $N = 1000$ nodes and a probability to generate one of the directed edges equal to $p = 0.002$.

In order to work with synthetic networks having an intrinsic community structure and a given reciprocity value, we use the benchmark generative model proposed in this paper. We generate networks with $N = 2100$ nodes and $K = 3$ communities by varying three different input parameters: the average degree $\langle k \rangle \in [2, 20]$, the reciprocity coefficient $\eta \in [0, 1)$ and the fraction of nodes with mixed membership $over \in [0, 1]$. While varying one of the parameter, the others are fixed to $\langle k \rangle = 20$, $\eta = 0.5$, and the degree of overlapping communities $over = 0$. In detail, networks are generated in two steps. First, membership vectors u and v are generated following an equal-size unmixed group membership and a Dirichlet distribution with parameter $\alpha = 0.1$ for the entries with mixed membership; and the affinity matrix w is generated using an assortative block structures with main probabilities $p_1 = K/N$ and secondary probabilities $p_2 = 0.1 p_1$. Thus the latent variables $\Theta = (u, v, w, \eta)$ are fixed. Second, edges are drawn according to the generative model described in the main text. Specifically, for each pair of nodes (i, j) , (i) extract A_{ij} from a Poisson of mean as in Eq. (10) and (ii) extract A_{ji} from a Poisson of mean as in Eq. (3). This procedure results in a directed network with the desired reciprocity and sparsity. We generate three independent networks for each value of the three different input parameters.

APPENDIX B: EDGE PREDICTION AND CROSS-VALIDATION

We perform edge prediction using 5-fold cross-validation. In each realization, we divide the dataset, i.e., the entries A_{ij} of the adjacency matrix, into five equal groups selected at random. We use four of these groups as a training set, to infer the parameters Θ . We then use the fifth group as a test set, evaluating the score for each A_{ij} in this set, and calculate the AUC value. By varying which group we use as the test set, we get five trials per realization. The final AUC is the average over these. To compute the regular AUC we use as score the expected value $\mathbb{E}_{P(A_{ij}|\Theta)}[A_{ij}] = m_{ij}$ as in Eq. (10); for the *conditional* AUC (AUC-cond), we use as score $\mathbb{E}_{P(A_{ij}|A_{ji}, \Theta)}[A_{ij}] = \lambda_{ij}^0 + \eta A_{ji}$, i.e., the expected value over the conditional distribution. Notice that the latter can only be computed for CRep, as for the others $m_{ij} \equiv \lambda_{ij}^0$, and thus the two AUC values coincide. The AUC is specified for binary entries, thus the edge weight is not accounted in the evaluation. However, our goal here is to assess edge existence, hence AUC is a suitable metric for this. If a practitioner aims at assessing the quality of the inferred weights as well, then one should specify different metrics for this.

APPENDIX C: INFERENCE: NUMERICAL IMPLEMENTATION

All the generative models require inferring K , the number of communities. We select this by cross-validation. Specifically, we run several held-out trials as explained above by varying K and select the value of K that gives the highest (regular) average AUC on the test sets. We then extract the

parameters of each method using their best K . For MT, BPF, and CRep₀, we extract the parameters u, v, w ; in addition, for CRep and CRep_{nc}, we extract η . All these algorithms converge to a local optima, as the likelihood landscape is not convex. Hence, we run the algorithm 10 times for different random initializations of the parameters and select the realization that has higher likelihood value.

APPENDIX D: DETAILED DERIVATIONS

We derive in detail the equations for inferring the parameters. We first apply a variational approach to make the problem tractable, and then use an expectation-maximization algorithm to derive the equations of the updates.

1. Variational approach

We aim at maximizing the log pseudolikelihood in Eq. (5). The first step is to facilitate the maximization process of the logarithmic term. We consider a probability distribution ρ_{ij} over the two competing terms: this is our estimate of the probability that the edges exist due to the contribution of either the community membership or the reciprocity term. Applying Jensen's inequality $\ln \bar{x} \geq \ln x$:

$$\begin{aligned} \ln \lambda_{ij} &= \ln \left(\rho_{ij}^{(1)} \frac{\lambda_{ij}^0}{\rho_{ij}^{(1)}} + \rho_{ij}^{(2)} \frac{\eta A_{ji}}{\rho_{ij}^{(2)}} \right) \\ &\geq \rho_{ij}^{(1)} \ln \frac{\lambda_{ij}^0}{\rho_{ij}^{(1)}} + \rho_{ij}^{(2)} \ln \frac{\eta A_{ji}}{\rho_{ij}^{(2)}} \\ &= \rho_{ij}^{(1)} \ln \sum_{k,q} u_{ik} v_{jq} w_{kq} + \rho_{ij}^{(2)} \ln \eta A_{ji} \\ &\quad - \rho_{ij}^{(1)} \ln \rho_{ij}^{(1)} - \rho_{ij}^{(2)} \ln \rho_{ij}^{(2)}. \end{aligned} \quad (D1)$$

Moreover, this holds with equality when

$$\rho_{ij}^{(1)} = \frac{\lambda_{ij}^0}{\lambda_{ij}^0 + \eta A_{ji}} \quad \text{and} \quad \rho_{ij}^{(2)} = \frac{\eta A_{ji}}{\lambda_{ij}^0 + \eta A_{ji}}. \quad (D2)$$

Thus maximizing $L^{ps}(\Theta)$ is equivalent to maximizing:

$$\begin{aligned} L^{ps}(\Theta, \rho) &= \sum_{i,j} \left\{ A_{ij} \left(\rho_{ij}^{(1)} \ln \sum_{k,q} u_{ik} v_{jq} w_{kq} + \rho_{ij}^{(2)} \ln \eta A_{ji} \right. \right. \\ &\quad \left. \left. - \rho_{ij}^{(1)} \ln \rho_{ij}^{(1)} - \rho_{ij}^{(2)} \ln \rho_{ij}^{(2)} \right) - \sum_{k,q} u_{ik} v_{jq} w_{kq} - \eta A_{ji} \right\}. \end{aligned}$$

We apply once more the variational approach to make the sum inside the logarithm tractable. Similarly as before, we introduce a probability distribution ϕ_{ijkq} such that

$$\ln \sum_{k,q} u_{ik} v_{jq} w_{kq} \geq \sum_{k,q} \phi_{ijkq} \ln u_{ik} v_{jq} w_{kq} - \sum_{k,q} \phi_{ijkq} \ln \phi_{ijkq}. \quad (D3)$$

The equality holds when

$$\phi_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k',q'} u_{ik'} v_{jq'} w_{k'q'}} = \frac{u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0}. \tag{D4}$$

Thus maximizing $L^{ps}(\Theta, \rho)$ is equivalent to maximizing:

$$L^{ps}(\Theta, \rho, \phi) = \sum_{i,j} \left\{ A_{ij} \rho_{ij}^{(1)} \left(\sum_{k,q} \phi_{ijkq} \ln u_{ik} v_{jq} w_{kq} - \sum_{k,q} \phi_{ijkq} \ln \phi_{ijkq} \right) + A_{ij} \rho_{ij}^{(2)} \ln \eta A_{ji} - A_{ij} \left(\rho_{ij}^{(1)} \ln \rho_{ij}^{(1)} + \rho_{ij}^{(2)} \ln \rho_{ij}^{(2)} \right) - \sum_{k,q} u_{ik} v_{jq} w_{kq} - \eta A_{ji} \right\} \tag{D5}$$

with respect to Θ, ρ, ϕ .

2. Expectation-Maximization updates

Equations for the updates of each of the parameters can be obtained by taking the derivative of Eq. (D5) with respect to a given parameter and setting it to zero. For instance, the update equation for η is obtained by considering the partial derivative:

$$\frac{\partial L^{ps}}{\partial \eta} = \sum_{i,j} \left[\frac{A_{ij} \rho_{ij}^{(2)}}{\eta} - A_{ji} \right]. \tag{D6}$$

Setting this to zero and defining $M = \sum_{i,j} A_{ij}$, we obtain

$$\eta = \frac{\sum_{i,j} A_{ij} \rho_{ij}^{(2)}}{\sum_{i,j} A_{ij}} = \frac{\eta}{M} \sum_{i,j} \frac{A_{ij} A_{ji}}{\lambda_{ij}}. \tag{D7}$$

Similarly, for the community affinity matrix, we get

$$w_{kq} = \frac{\sum_{i,j} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}}{\sum_{i,j} u_{ik} v_{jq}}. \tag{D8}$$

Here we show how to enforce constraints like $\sum_k u_{ik} = 1$, which is an arbitrary choice that can be easily incorporated into our model. To this end, it is convenient to rewrite the log pseudolikelihood as follow:

$$L^{ps}(\Theta, \rho, \phi) = F(u_{ik}, v_{jq}, w_{kq}) - \sum_{i,j,k,q} u_{ik} v_{jq} w_{kq}, \tag{D9}$$

Then, following the approach in Ref. [53], to simplify the maximization of the log pseudolikelihood, we substitute w_{kq} from Eq. (D8) into Eq. (D9):

$$\begin{aligned} L^{ps}(\Theta, \rho, \phi) &= F(u_{ik}, v_{jq}, w_{kq}) \\ &\quad - \sum_{i,j,k,q} u_{ik} v_{jq} \frac{\sum_{i,j} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}}{\sum_{i,j} u_{ik} v_{jq}} \\ &= F(u_{ik}, v_{jq}, w_{kq}) \\ &\quad - \sum_{k,q} \sum_{i,j} u_{ik} v_{jq} \frac{\sum_{i,j} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}}{\sum_{i,j} u_{ik} v_{jq}} \\ &= F(u_{ik}, v_{jq}, w_{kq}) - \sum_{i,j,k,q} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}. \end{aligned} \tag{D10}$$

The second term in the above equation does not depend explicitly on u_{ik} and v_{jq} . In order to apply the constraint on the maximization, we add Lagrange multipliers γ_i^u, γ_i^v :

$$\begin{aligned} L^{ps}(\Theta, \rho, \phi) &= F(u_{ik}, v_{jq}, w_{kq}) \\ &\quad - \sum_{k,q} \sum_{i,j} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq} - \gamma_i^u \left(\sum_k u_{ik} - 1 \right) \\ &\quad - \gamma_j^v \left(\sum_q v_{jq} - 1 \right). \end{aligned} \tag{D11}$$

The update equation for u_{ik} is obtained by considering the partial derivative

$$\frac{\partial L^{ps}}{\partial u_{ik}} = \sum_{j,q} \left(\frac{A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}}{u_{ik}} \right) - \gamma_i^u, \tag{D12}$$

and setting it to zero, which yields

$$u_{ik} = \frac{1}{\gamma_i^u} \sum_{j,q} A_{ij} \rho_{ij}^{(1)} \phi_{ijkq}. \tag{D13}$$

By applying the normalization constraint on the u_{ik} , i.e., $\sum_k u_{ik} = 1$, and noticing that $\rho_{ij}^{(1)} \phi_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0 + \eta A_{ji}}$, we can find an expression for γ_i^u :

$$\gamma_i^u = \sum_{j,k,q} \frac{A_{ij} u_{ik} v_{jq} w_{kq}}{\lambda_{ij}^0 + \eta A_{ji}} = \sum_j \frac{A_{ij} \lambda_{ij}^0}{\lambda_{ij}^0 + \eta A_{ji}}. \tag{D14}$$

Similarly, we have the following update equation for v :

$$v_{jq} = \frac{1}{\gamma_j^v} \sum_{i,k} A_{ji} \rho_{ij}^{(1)} \phi_{jikq}, \tag{D15}$$

where

$$\gamma_j^v = \sum_{i,k,q} \frac{A_{ji} u_{ik} v_{jq} w_{kq}}{\lambda_{ji}^0 + \eta A_{ij}} = \sum_i \frac{A_{ji} \lambda_{ji}^0}{\lambda_{ji}^0 + \eta A_{ij}}. \tag{D16}$$

3. Deriving the expected value of the marginal distribution

$$\begin{aligned} \mathbb{E}[A_{ij}] &= m_{ij} = \sum_{A_{ij}, A_{ji}} A_{ij} P(A_{ij}, A_{ji} | \Theta) \\ &= \sum_{A_{ji}} P(A_{ji} | \Theta) \sum_{A_{ij}} A_{ij} P(A_{ij} | A_{ji}, \Theta) \end{aligned}$$

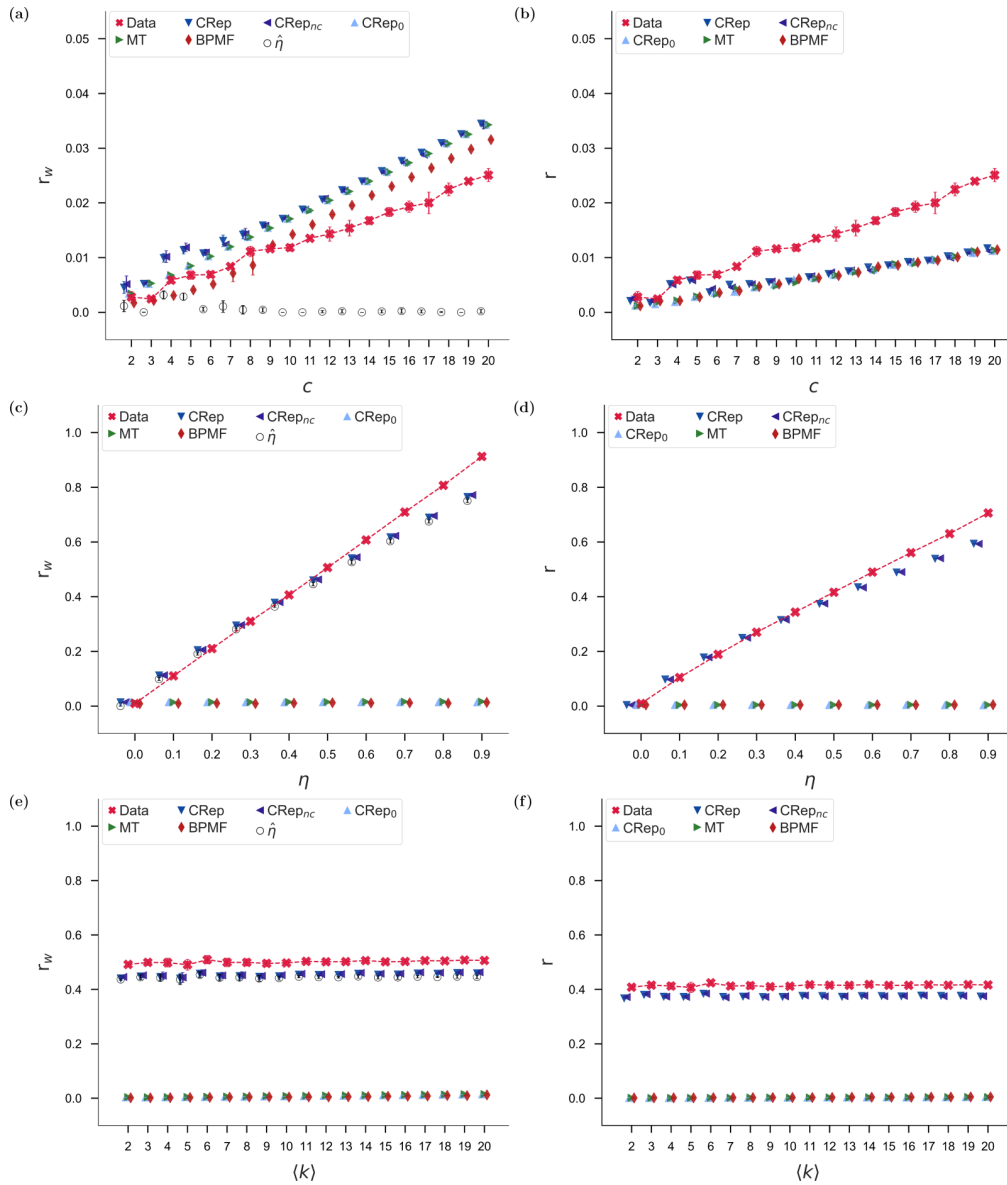


FIG. 8. Reciprocity in synthetic networks. Synthetic networks with $N = 2100$ nodes and $K = 3$ communities of equal-size unmixed group membership generated with a stochastic block model [(a) and (b)] by varying the average degree within the same community c and our benchmark generative model, by varying the reciprocity parameter η [(c) and (d)] and the average degree (k) [(e) and (f)]. Results are empirical averages and standard deviations over 15 samples of three independent synthetic networks (five sample per input network). The red markers indicate the average on the three input networks. [(a), (c), and (e)] The quantity r_w as defined in Eq. (15); $\hat{\eta}$ is the inferred parameter in CRep and CRep_{nc}. [(b), (d), and (f)] Standard reciprocity r .

$$\begin{aligned}
 &= \sum_{A_{ji}} P(A_{ji}|\Theta) [\lambda_{ij}^0 + \eta A_{ji}] \\
 &= \lambda_{ij}^0 + \eta \sum_{A_{ji}} A_{ji} P(A_{ji}|\Theta) \\
 &= \lambda_{ij}^0 + \eta m_{ji} \\
 &= \lambda_{ij}^0 + \eta (\lambda_{ji}^0 + \eta m_{ij}). \tag{D17}
 \end{aligned}$$

Solving for m_{ij} yields

$$m_{ij} (1 - \eta^2) = (\lambda_{ij}^0 + \eta \lambda_{ji}^0), \tag{D18}$$

which implies

$$m_{ij} = \frac{\lambda_{ij}^0 + \eta \lambda_{ji}^0}{(1 - \eta^2)}. \tag{D19}$$

4. Expected value of \mathbf{r}_w

With similar calculations as before we obtain

$$\mathbb{E}[A_{ij} A_{ji}] = \sum_{A_{ij}, A_{ji}} A_{ij} A_{ji} P(A_{ij}, A_{ji}|\Theta) \tag{D20}$$

$$= \lambda_{ij}^0 m_{ji} + \eta \mathbb{E}[A_{ji}^2]. \tag{D21}$$

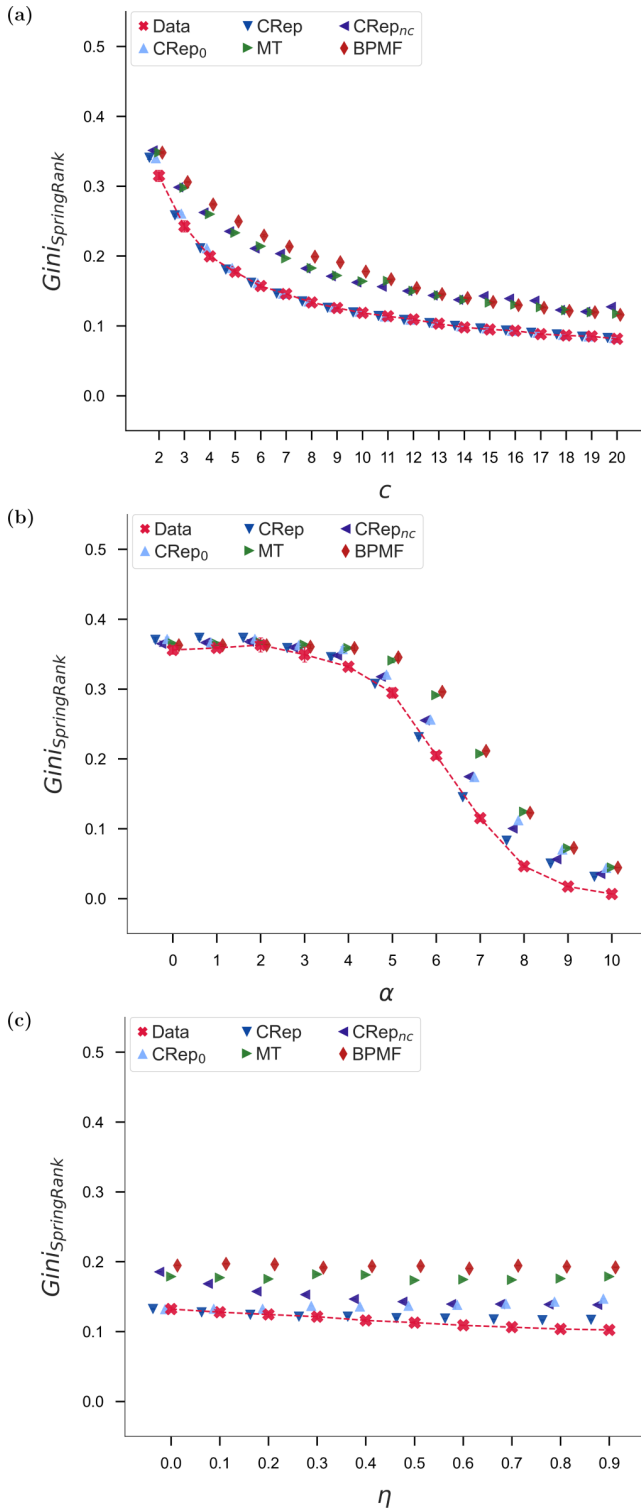


FIG. 9. Hierarchical structure in synthetic networks. Synthetic networks generated with (a) the stochastic block model, (b) the HL model, and (c) the benchmark generative model. Results are averages and standard deviations of the Gini index on SPRINGRANK ranking scores over 15 samples of three independent synthetic networks (five sample per input network). The red markers indicate the average on the three input networks.

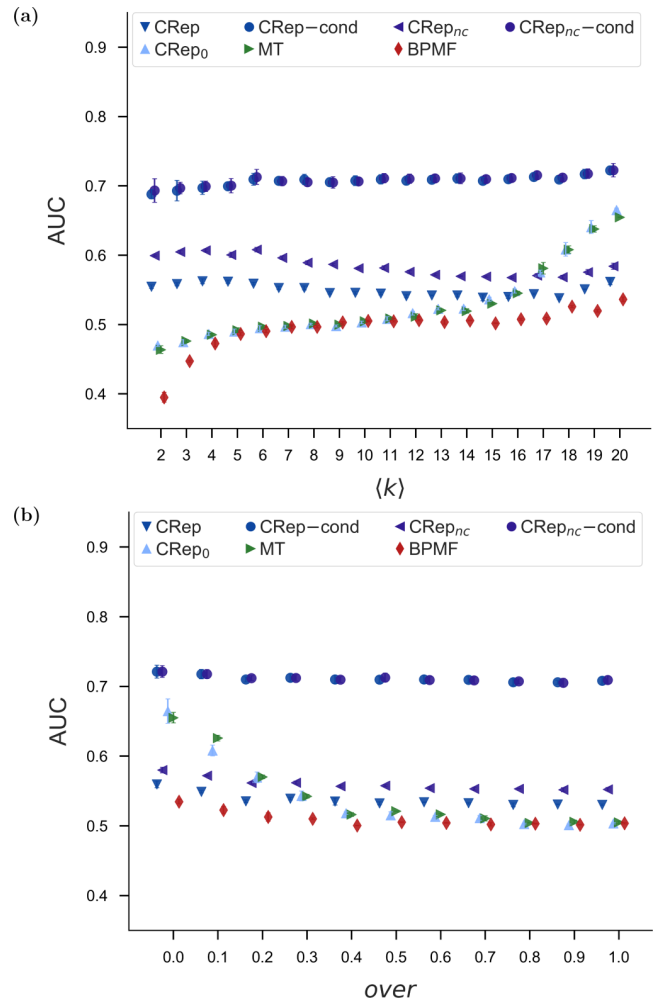


FIG. 10. Edge prediction in synthetic networks. Synthetic networks with $N = 2100$ nodes and $K = 3$ communities of equal-size unmixed group membership generated with the benchmark generative model proposed above by varying (a) the average degree $\langle k \rangle$ and (b) the fraction of nodes with mixed membership *over*. The results are averages and standard deviations over three independent synthetic networks and over 5-fold cross-validation test sets. The accuracy of edge prediction is measured with AUC and the baseline is the random value 0.5.

To fully determine this expression we need to specify the second moment $\mathbb{E}[A_{ji}^2]$. For binary variables, we could assume $\mathbb{E}[A_{ji}^2] = \mathbb{E}[A_{ji}] = m_{ji}$, as this is the case for Bernoulli distributions. With this assumption, we obtain $\mathbb{E}[A_{ij} A_{ji}] = (\lambda_{ij}^0 + \eta)m_{ji}$. Alternatively, we can assume $\mathbb{E}[A_{ji}^2] = m_{ji} + m_{ji}^2$ as is the case for the Poisson distribution, and thus obtain $\mathbb{E}[A_{ij} A_{ji}] = (\lambda_{ij}^0 + \eta)m_{ji} + \eta m_{ji}^2$. Finally we have

$$\begin{aligned} \mathbb{E}[r_w] &= \mathbb{E}\left[\frac{\sum_{i,j} [A_{ij} A_{ji}]}{\sum_{i,j} [A_{ij}]}\right] \approx \frac{\sum_{i,j} \mathbb{E}[A_{ij} A_{ji}]}{\sum_{i,j} \mathbb{E}[A_{ij}]} \\ &= \frac{\sum_{i,j} [(\lambda_{ij}^0 + \eta)m_{ji} + \eta m_{ji}^2]}{\sum_{i,j} m_{ij}} \\ &= \eta + \frac{\sum_{i,j} [\lambda_{ij}^0 m_{ji} + \eta m_{ji}^2]}{\sum_{i,j} m_{ij}} \geq \eta, \end{aligned} \tag{D22}$$

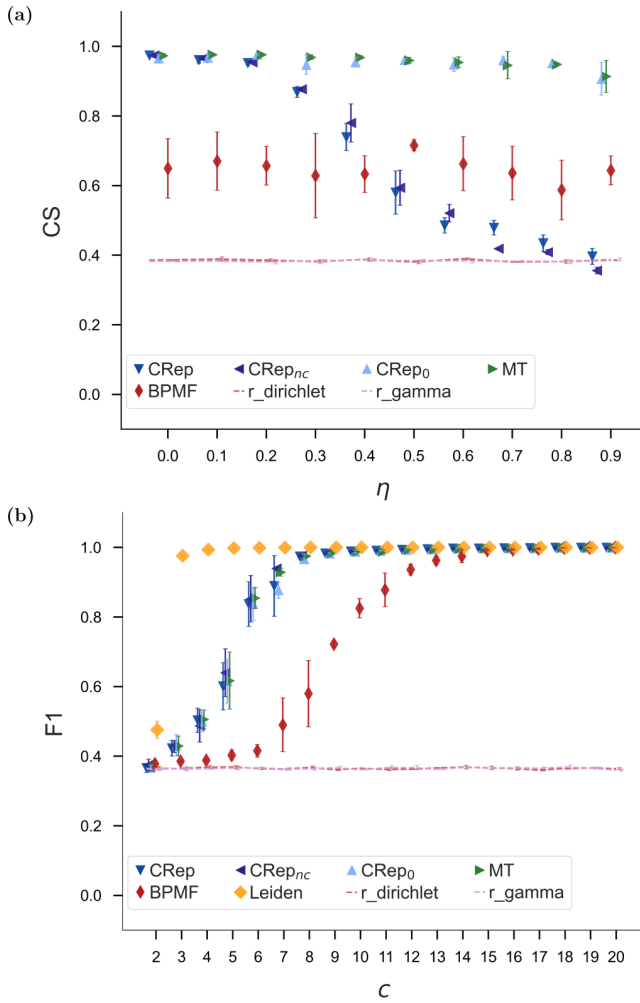


FIG. 11. Community detection in synthetic networks. Synthetic networks with $N = 2100$ nodes and $K = 3$ communities of equal-size unmixed group membership generated with (a) the benchmark generative model proposed above by varying the reciprocity parameter η , and (b) a stochastic block model. The results are averages and standard deviations over three independent synthetic networks. The accuracy of community detection is measured with (a) cosine similarity and (b) with F1-score as similarity measure, and values close to 1 means higher similarity. The dashed lines represent random baselines, where membership u_i are extracted randomly from a Dirichlet of parameter $\alpha = 0.1$ or a Gamma distribution of parameters $\alpha = 0.1$ and $\beta = 1$, to enforce sparsity.

where in the first row we use the first-order Taylor expansion as an approximation. With this assumption, we obtain that the parameter η is a lower bound for the expected value of r_w . An equivalent expression can be derived for models that assume conditional independence, e.g., our model with $\eta = 0$. In this case, we get

$$\begin{aligned} \mathbb{E}[A_{ij} A_{ji}] &= \sum_{A_{ij}} A_{ij} P(A_{ij}|\Theta) \sum_{A_{ji}} A_{ji} P(A_{ji}|\Theta) \\ &= m_{ij} m_{ji}, \end{aligned} \quad (\text{D23})$$

which yields

$$\mathbb{E}[r_w] = \frac{\sum_{i,j} \mathbb{E}[A_{ij} A_{ji}]}{\sum_{i,j} \mathbb{E}[A_{ij}]} = \frac{\sum_{i,j} m_{ij} m_{ji}}{\sum_{i,j} m_{ij}}. \quad (\text{D24})$$

APPENDIX E: HOLLAND AND LEINHARDT RECIPROCALITY MODEL

The model assumes an unweighted and directed network, i.e., asymmetric adjacency matrix with binary values $A_{ij} \in \{0, 1\}$, and the following joint probability:

$$P(A|\theta, \alpha) = \frac{e^{-H(A, \theta, \alpha)}}{Z(\theta, \alpha)^{\frac{n(n-1)}{2}}}, \quad (\text{E1})$$

$$H(A, \theta, \alpha) = \theta \sum_{i < j} (A_{ij} + A_{ji}) - \alpha \sum_{i < j} A_{ij} A_{ji}, \quad (\text{E2})$$

where $Z(\theta, \alpha) = 1 + 2e^{-\theta} + e^{-2\theta + \alpha}$ is the normalization term. The parameter α controls the level of reciprocity, it couples the two entries A_{ij} and A_{ji} thus making the model not factorized; edges between different pairs (i, j) are conditionally independent given the parameters. This is one of the few analytically tractable exponential random graph models. Due to this property, we can extract analytical marginal and conditional distributions for a pair of nodes (i, j) :

$$P(A_{ij}|\theta, \alpha) = \frac{e^{-\theta A_{ij}} + e^{-\theta - A_{ij}(\theta - \alpha)}}{Z(\theta, \alpha)}, \quad (\text{E3})$$

$$P(A_{ji}|A_{ij}, \theta, \alpha) = \frac{e^{-A_{ji}(\theta - \alpha A_{ij})}}{1 + e^{-(\theta - \alpha A_{ij})}}. \quad (\text{E4})$$

These expressions can be used to sample networks with the joint distribution given in Eq. (E2). Tuning the value of the parameter α , one generates networks with different values of reciprocity.

APPENDIX F: PERFORMANCE IN SYNTHETIC NETWORKS

1. Reproducing the topological properties

Here we show in more details the ability of the models to reproduce network samples that replicate relevant network quantities. Figure 8 shows r and r_w as defined in Eq. (15), computed in the sampled networks of synthetic data generated with a stochastic block model and our benchmark generative model. As expected, the reciprocity in networks generated with the stochastic block model is always close to zero. Instead, the networks generated with our benchmark generative model present different values of reciprocity, and CRRep captures these values significantly better than the other models, consistently across various magnitudes of input η . Even in the case of fixed η , by changing sparsity, we observe the same pattern. By varying the degree of overlapping communities, we obtain the same results as changing the average degree (we do not report them here).

Figure 9 shows the Gini index computed on nodes scores obtained with the SPRINGRANK algorithm. The Gini index provides a global measure for the whole network, the higher its value, the more hierarchical the network is. We compare the average over the five samples, and we find that CRRep and CRRep₀ have reasonable accuracy in retrieving the Gini

TABLE I. Datasets description.

Network	Abbreviation	Category	N	E	Ref.
Dutch college	DT2	Human Social Network	26	144	[54]
Dutch college	DT6	Human Social Network	30	256	[54]
Highschool Friendships	HST11	Human Social Network	31	100	[54]
Highschool Friendships	HST12	Human Social Network	30	114	[54]
Highschool Friendships	HST2	Human Social Network	62	245	[54]
Online dating	POK0	Human Social Network	3562	18098	[55]
Online dating	POK6	Human Social Network	3227	10696	[55]
Online dating	POK12	Human Social Network	2530	7653	[55]
Physicians	Phys	Human Social Network	95	458	[54]
Seventh graders	7th	Human Social Network	29	376	[54]
Adolescent health	AdH	Human Social Network	2213	11676	[54]
Advogato	Adv	Online Social network	3858	42188	[54]
Faculty hiring, business department	BS	Institutions Social Network	112	3321	[56]
Faculty hiring, computer department	CS	Institutions Social Network	198	2702	[56]
Faculty hiring, history department	HS	Institutions Social Network	140	2242	[56]
Erasmus Mobility Statistics 2014	ERs14	Institutions Social Network	2264	79532	[34]
Erasmus Mobility Statistics 2015	ERs15	Institutions Social Network	2890	79665	[34]
Erasmus Mobility Statistics 2016	ERs16	Institutions Social Network	3713	85468	[34]
Erasmus Mobility Statistics 2017	ERs17	Institutions Social Network	4200	89792	[34]
Erasmus Mobility Statistics 2018	ERs18	Institutions Social Network	4389	90972	[34]
Citation 2005	CIT05	Citation Network	2130	11153	[57]
Statistics Citation	SCC2016	Citation Network	2654	21568	[58]
ACM v9 2012	ACMv9	Citation Network	8469	56801	[59]
Email Eu core network	EU	Email Network	834	24348	[57]
DNC Email	DNC	Email Network	548	3575	[54]
Wiki Talk ht	Wiki	Communication Network	80	164	[54]
UC Social	UCS	Communication Network	1302	19044	[54]
Blogs	Blg	Hyperlink Network	830	16107	[54]
Cattle	Ctl	Animal Network	24	191	[54]
FAA Preferred Routes	FAA	Infrastructure Network	1064	2275	[54]

index of the original network, while the other models tend to overestimate it. This is consistent over the various synthetic network topologies, i.e., network generated with the stochastic block model, panel (a), the HL model, panel (b), and our benchmark generative model, panel (c). Furthermore, we notice that this topological property decreases as the average degree within the same community, c , and α increase, while it is not influenced by the value of η . We omit the results for the networks generated with our benchmark generative model by varying the sparsity and the fraction of nodes with

mixed-membership because we obtain similar results to the stochastic block networks and the benchmark data by varying η , respectively.

2. Edge prediction in synthetic networks

Here we show the results in terms of edge prediction on synthetic data generated with our benchmark generative model by varying the average degree $\langle k \rangle$ and the fraction of nodes with mixed membership, which we denote *over*.

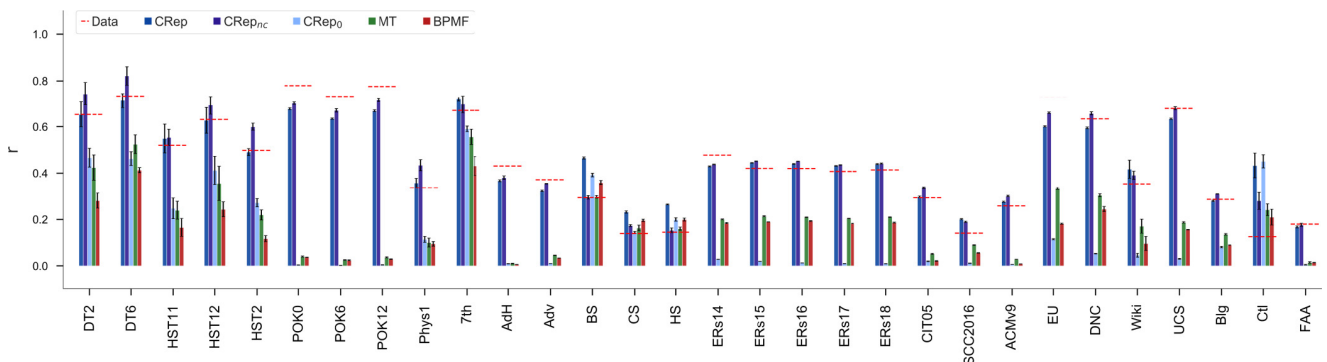


FIG. 12. Reciprocity in real networks. Empirical averages and standard deviations of reciprocity r over five samples of each real network (see Table I for details). The red dashed lines indicate the r on the input networks.

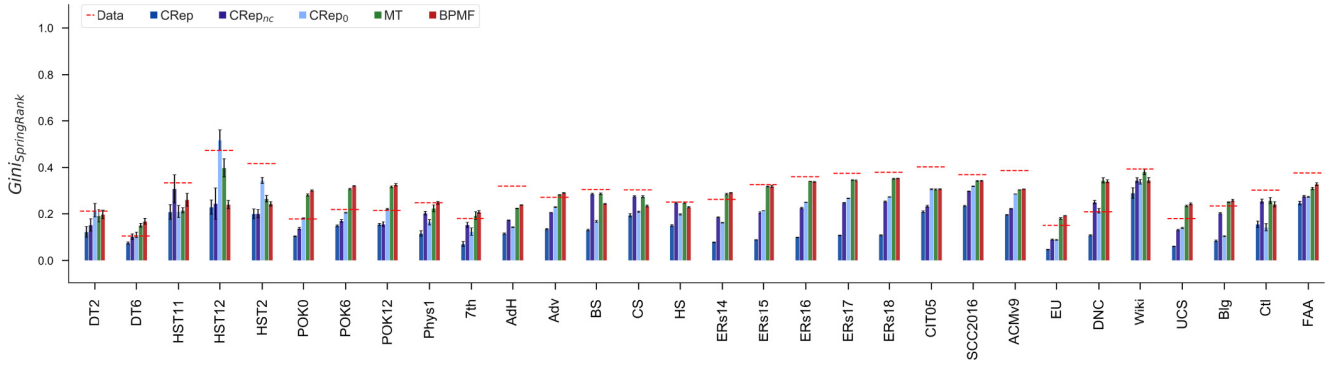


FIG. 13. Hierarchical structure in real networks. Empirical averages and standard deviations of the Gini index on SpringRank ranking scores over five samples of each real network (see Table I for details). The red dashed lines indicate the values on the input networks.

We use both conditional and regular edge prediction and Figure 10 highlights the robustness of CRep and CRep_{nc} in terms of conditional edge predictions, as their performance are significantly higher than that of the other algorithms and

do not decrease with increasing overlapping communities and sparsity. Indeed, the results are robust, as we vary the fraction of nodes with overlapping community membership and the average degree, while fixing $\eta = 0.5$. Notice also the stability

TABLE II. Edge prediction in real networks. Regular AUC and conditional AUC (AUC-cond) for all real networks (see Table I for details). Results are averages and standard deviations over 5-fold cross-validation test sets. In grey box, we show the best performance over all methods, while in boldface the best results in terms of regular AUC. The last row reports the average and standard deviation of each method over datasets.

Dataset	AUC						AUC-cond	
	CRep	CRep _{nc}	CRep ₀	MT	BMPF	OLP	CRep	CRep _{nc}
DT2	0.71 ± 0.01	0.73 ± 0.01	0.653 ± 0.009	0.71 ± 0.03	0.72 ± 0.01	0.712	0.77 ± 0.02	0.79 ± 0.03
DT6	0.72 ± 0.03	0.76 ± 0.01	0.72 ± 0.01	0.762 ± 0.006	0.774 ± 0.008	0.737	0.83 ± 0.03	0.85 ± 0.02
HST11	0.74 ± 0.01	0.73 ± 0.01	0.63 ± 0.03	0.62 ± 0.03	0.63 ± 0.04	0.714	0.78 ± 0.02	0.76 ± 0.02
HST12	0.82 ± 0.02	0.801 ± 0.008	0.743 ± 0.004	0.74 ± 0.01	0.76 ± 0.02	0.778	0.85 ± 0.01	0.86 ± 0.02
HST2	0.771 ± 0.009	0.76 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.71 ± 0.01	0.828	0.808 ± 0.009	0.79 ± 0.02
POK0	0.7747 ± 0.0001	0.845 ± 0.002	0.665 ± 0.002	0.7400 ± 0.0009	0.7652 ± 0.0002	0.804	0.908 ± 0.002	0.934 ± 0.002
POK6	0.758 ± 0.001	0.818 ± 0.002	0.587 ± 0.003	0.626 ± 0.002	0.6939 ± 0.0007	0.750	0.884 ± 0.005	0.909 ± 0.002
POK12	0.765 ± 0.002	0.833 ± 0.002	0.582 ± 0.002	0.606 ± 0.002	0.6723 ± 0.0006	0.739	0.905 ± 0.003	0.924 ± 0.002
Phys1	0.600 ± 0.008	0.627 ± 0.006	0.556 ± 0.009	0.57 ± 0.01	0.60 ± 0.02	0.577	0.676 ± 0.005	0.71 ± 0.01
7th	0.69 ± 0.02	0.79 ± 0.01	0.72 ± 0.02	0.800 ± 0.009	0.809 ± 0.005	0.494	0.77 ± 0.01	0.84 ± 0.01
AdH	0.678 ± 0.003	0.696 ± 0.002	0.656 ± 0.002	0.666 ± 0.003	0.627 ± 0.004	0.867	0.760 ± 0.003	0.787 ± 0.001
Adv	0.771 ± 0.002	0.8919 ± 0.0001	0.760 ± 0.003	0.887 ± 0.001	0.8907 ± 0.0005	0.940	0.830 ± 0.002	0.9333 ± 0.0005
BS	0.662 ± 0.004	0.8749 ± 0.0006	0.649 ± 0.004	0.8749 ± 0.0005	0.8746 ± 0.0009	0.711	0.66 ± 0.01	0.8750 ± 0.0006
CS	0.715 ± 0.008	0.829 ± 0.001	0.696 ± 0.005	0.830 ± 0.002	0.838 ± 0.001	0.844	0.709 ± 0.008	0.833 ± 0.001
HS	0.661 ± 0.005	0.866 ± 0.003	0.646 ± 0.003	0.866 ± 0.003	0.872 ± 0.001	0.865	0.654 ± 0.005	0.867 ± 0.003
ERs14	0.754 ± 0.001	0.9157 ± 0.0005	0.696 ± 0.009	0.9115 ± 0.0004	0.9123 ± 0.0003	0.893	0.810 ± 0.001	0.9278 ± 0.0002
ERs15	0.79 ± 0.01	0.9361 ± 0.0002	0.72 ± 0.02	0.9330 ± 0.0002	0.9312 ± 0.0002	0.929	0.82 ± 0.01	0.9454 ± 0.0002
ERs16	0.8057 ± 0.0006	0.9454 ± 0.0002	0.7064 ± 0.0004	0.9402 ± 0.0003	0.9419 ± 0.0001	0.944	0.8346 ± 0.0006	0.9552 ± 0.0002
ERs17	0.822 ± 0.005	0.9484 ± 0.0001	0.734 ± 0.002	0.9433 ± 0.0002	0.9468 ± 0.0002	0.950	0.838 ± 0.005	0.9568 ± 0.0002
ERs18	0.8334 ± 0.0006	0.9501 ± 0.0001	0.732 ± 0.002	0.9444 ± 0.0002	0.9490 ± 0.0002	0.952	0.8476 ± 0.0006	0.9579 ± 0.0001
CIT05	0.910 ± 0.002	0.9189 ± 0.0008	0.901 ± 0.001	0.918 ± 0.001	0.908 ± 0.001	0.954	0.928 ± 0.002	0.9389 ± 0.0008
SCC2016	0.893 ± 0.001	0.923 ± 0.001	0.8938 ± 0.0009	0.925 ± 0.001	0.9211 ± 0.0007	0.946	0.901 ± 0.001	0.925 ± 0.001
ACMv9	0.926 ± 0.001	0.9350 ± 0.0007	0.919 ± 0.001	0.9352 ± 0.0001	0.9254 ± 0.0006	0.968	0.941 ± 0.001	0.9525 ± 0.0007
EU	0.795 ± 0.007	0.9297 ± 0.0004	0.760 ± 0.007	0.9264 ± 0.0008	0.9169 ± 0.0006	0.944	0.926 ± 0.007	0.9619 ± 0.0006
DNC	0.766 ± 0.003	0.929 ± 0.002	0.730 ± 0.001	0.8566 ± 0.0003	0.913 ± 0.001	0.919	0.890 ± 0.006	0.939 ± 0.002
Wiki	0.68 ± 0.02	0.70 ± 0.02	0.63 ± 0.01	0.63 ± 0.02	0.83 ± 0.01	0.801	0.73 ± 0.01	0.76 ± 0.02
UCS	0.754 ± 0.005	0.8762 ± 0.0008	0.717 ± 0.003	0.8558 ± 0.0008	0.844 ± 0.002	0.850	0.904 ± 0.005	0.9530 ± 0.0008
Blg	0.784 ± 0.001	0.9312 ± 0.0001	0.767 ± 0.002	0.9321 ± 0.0003	0.9334 ± 0.0001	0.924	0.824 ± 0.001	0.9463 ± 0.0001
Ctl	0.56 ± 0.03	0.66 ± 0.02	0.57 ± 0.03	0.67 ± 0.02	0.70 ± 0.03	0.574	0.56 ± 0.03	0.66 ± 0.02
FAA	0.576 ± 0.003	0.589 ± 0.002	0.543 ± 0.007	0.535 ± 0.004	0.607 ± 0.003	0.779	0.592 ± 0.002	0.595 ± 0.002
Avg.	0.749 ± 0.007	0.831 ± 0.004	0.700 ± 0.007	0.796 ± 0.006	0.813 ± 0.006	0.823	0.804 ± 0.005	0.867 ± 0.006

TABLE III. Extended features used in the link prediction process for a directed network.

Feature	Description
Common neighbors out/in	defined for a pair of nodes: $x, y: \Gamma(x)_{out/in} \cap \Gamma(y)_{out/in} $
Jaccard index	defined for a pair of nodes: $x, y: \frac{ \Gamma(x)_{out/in} \cap \Gamma(y)_{out/in} }{ \Gamma(x)_{out/in} \cup \Gamma(y)_{out/in} }$
Adamic-Adar index	defined for a pair of nodes: $x, y: \sum_{z \in \{\Gamma(x)_{out/in} \cap \Gamma(y)_{out/in}\}} \frac{1}{\ln \Gamma(z) }$
Resource Allocation index	defined for a pair of nodes: $x, y: \sum_{z \in \{\Gamma(x)_{out/in} \cap \Gamma(y)_{out/in}\}} \frac{1}{ \Gamma(z) }$
Betweenness centrality	a measure of node centrality based on the shortest paths
Closeness centrality	defined for a pair of nodes: $x, y: \frac{1}{\sum_y d(y,x)}$
Shortest Paths	shortest path between nodes: x, y
Katz centralities	a measure of centrality in a network
PageRank centralities	a measure of the importance of a node as an adjustment of Katz centrality
Eigenvector centralities	an adjustment of Katz centrality of a node in regards to the importance of its neighbors
Clustering coefficient for node x	$\frac{\text{number of triangles connected to node } x}{\text{number of triples centered around node } x}$
Preferential attachment	the tendency of nodes to connect to the nodes with higher degree
Common community	1 if the pair of nodes belong to the same community, otherwise zero

of CRep and $CRep_{nc}$ in terms of regular edge prediction and how they outperform the other models in critical ranges, e.g., small $\langle k \rangle$ and high *over*.

Moreover, we find more stable results also in terms of regular edge prediction, where CRep and $CRep_{nc}$ have constant values across the different input parameters, outperforming other methods in critical ranges, e.g., small average degree or high overlap between communities. The results of our experiments suggest that working with conditional probabilities results in more robust estimates of the probability that an edge exists if we have access to the edge in the opposite direction. Performance improvement is more significant when community structure is not the predominant mechanism in edge formation.

3. Community detection in synthetic data

For sake of completeness, here we show the performance of the models on recovering communities. We consider as performance measure the F1-score (F1) and cosine similarity (CS), the former one is valid for hard membership while the latter captures mixed-membership, we calculate for both the average over the nodes. When measuring the F1-score we consider the entries of maximum value of the membership vectors. Both measures are between 0 and 1 and a value of 1 means perfect reconstruction. Figure 11 shows the accuracy in networks generated with the benchmark generative model by varying the reciprocity parameter η and for synthetic data created with a stochastic block model by varying the average degree within the same community c . For comparison in these last networks, we consider also the Leiden algorithm [60], a nongenerative method. Even if community detection is not the main focus of our model, we notice the ability of CRep in retrieving communities in networks without reciprocity, while its performance decreases as reciprocity increases. This is expected as the community impact in determining the likelihood of an edge decreases as η increases. Notice that the benchmark data have been generated with fixed $\langle k \rangle = 20$, thus models without reciprocity are capable of fully recovering

the community even in the case where reciprocity is there, provided that the average degree is large enough. These synthetic tests suggest, on one side, the robustness of community detection-only methods in recovering communities even in the presence of reciprocity; on the other side the good performance of CRep in recovering communities when reciprocity has intermediate or low level. This is somehow expected, as this model gives increasingly less weight to the communities as reciprocity increases, thus it is not optimized to recover the communities when these are not fully determining edge formation.

APPENDIX G: PERFORMANCE IN REAL NETWORKS

1. Real data: dataset description

We apply our approach to different types of networks, such as social, infrastructure, online communication, and citation networks. Table I provides a brief overview of the datasets studied in this work, as well as their abbreviations. All datasets, have been pre-processed as follows: (i) self-loops are removed; (ii) only nodes that have at least one out-going and one in-coming edge are kept; (iii) we used only the giant connected components. Some datasets require additional specific preprocessing. Specifically, the citation networks (here CIT05, SCC2016, ACMv9) require extracting a network author-author from a network of paper-citation, so that an edge means that an author cites another author. Furthermore, we split dynamic networks into separate individual networks where we kept only interactions happening within a certain time window. This applies to Dutch (DT2, DT6), High school friendships (HST11, HST12, HST2), online dating (POK0, POK6, POK12), and Erasmus (ERs14, ERs15, ERs16, ERs17, ERs18).

2. Reproducing the topological properties

Here we show the ability of the models to reproduce network samples that replicate relevant network quantities. For each real network, we infer the parameters by each model,

and use them to generate five synthetic network samples. Figure 12 shows the reciprocity r . For each model, it outputs the averages and the standard deviations over the five samples and the dashed red lines indicate the r value of the input datasets. We notice the heterogeneity of the analysed networks and how CRep adapts to all different situations, while the other models underestimate the true value most of the times.

Figure 13 shows the Gini index computed on nodes scores obtained with the SPRINGRANK algorithm. The results vary widely depending on the datasets, and we cannot draw general conclusions. In this scenario, we have also studied the reproducibility of the clustering coefficient, i.e., the tendency of nodes to form edges within the same neighborhood, however, we obtain poor results in line with the SBM approach, as predicted in Ref. [46]. Moreover, these are topological properties that involve more complex interactions than pairwise, as in the case of reciprocity (clustering involves triangles and SPRINGRANK score is a global measure). This suggests that, in order to have better performance, one would need to develop more complex models, for instance extending the ideas behind CRep to capture triadic interactions, possibly guided by domain-knowledge about how triadic interactions and reciprocity are related [45]. We leave this for future work, noting that while exponential random graph models can do this, they do not include latent community structure (analogously as for reciprocity).

3. Link prediction features

Here we present the supervised learning-link prediction routine (OLP) used for comparison in the edge prediction task on real data. In the link prediction task, scores are assigned to all possible pairs of nodes in the graph based on a set of criteria. Then, the pairs of nodes are sorted according to their scores in an ascending order and the most-likely links are the pairs with scores above a threshold value.

Two categories of features are used to determine the criteria of link classification: (i) global features, defined based on the features of the entire network, such as the number of nodes, number of edges, average degree of nodes, and the average clustering coefficient, and (ii) local features, which include the descriptive features of a single node or a pair of nodes.

In this work, we apply the extended definition of features for a directed network of Ghasemian *et al.* [32]. We also examine the effect of belonging to the same community on the local pairwise features, i.e., pairwise attributes contribute in the link prediction only if the two nodes belong to the same community. However, we did not find significant changes and at the price of higher computational cost, hence, we exclude this factor from the study and omit the results. Considering $\Gamma(x)_{out/in}$ as the set of out/in-neighbors of node x , and $d(x, y)$ as the distance between nodes x and y , some of the well-known features deployed for link prediction are presented in Table III.

-
- [1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).
 - [2] L. D. Molm, The structure of reciprocity, *Soc. Psychol. Quart.* **73**, 119 (2010).
 - [3] M. A. Nowak and K. Sigmund, Evolution of indirect reciprocity, *Nature (London)* **437**, 1291 (2005).
 - [4] D. Garlaschelli and M. I. Loffredo, Structure and evolution of the world trade network, *Physica A* **355**, 138 (2005).
 - [5] K. Zhao, X. Wang, M. Yu, and B. Gao, User recommendations in reciprocal and bipartite social networks—an online dating case study, *IEEE Intell. Syst.* **29**, 27 (2013).
 - [6] J. Wincent, S. Anokhin, D. Örtqvist, and E. Autio, Quality meets structure: Generalized reciprocity and firm-level advantage in strategic networks, *J. Manage. Stud.* **47**, 597 (2010).
 - [7] W. Li, T. Aste, F. Caccioli, and G. Livan, Reciprocity and impact in academic careers, *EPJ Data Sci.* **8**, 20 (2019).
 - [8] M. E. J. Newman, S. Forrest, and J. Balthrop, Email networks and the spread of computer viruses, *Phys. Rev. E* **66**, 035101(R) (2002).
 - [9] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic block-models: First steps, *Soc. Netw.* **5**, 109 (1983).
 - [10] P. W. Holland and S. Leinhardt, An exponential family of probability distributions for directed graphs, *J. Am. Stat. Assoc.* **76**, 33 (1981).
 - [11] J. Park and M. E. J. Newman, Statistical mechanics of networks, *Phys. Rev. E* **70**, 066117 (2004).
 - [12] D. Garlaschelli and M. I. Loffredo, Multispecies grand-canonical models for networks with reciprocity, *Phys. Rev. E* **73**, 015101(R) (2006).
 - [13] S. Wasserman and C. Anderson, Stochastic a posteriori block-models: Construction and assessment, *Soc. Netw.* **9**, 1 (1987).
 - [14] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, An introduction to exponential random graph (p*) models for social networks, *Soc. Netw.* **29**, 173 (2007).
 - [15] T. Squartini, F. Picciolo, F. Ruzzenenti, and D. Garlaschelli, Reciprocity of weighted networks, *Sci. Rep.* **3**, 2729 (2013).
 - [16] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks, *Phys. Rev. E* **95**, 042317 (2017).
 - [17] B. Ball, B. Karrer, and M. E. J. Newman, Efficient and principled method for detecting communities in networks, *Phys. Rev. E* **84**, 036103 (2011).
 - [18] J. R. Lloyd, P. Orbanz, Z. Ghahramani, and D. M. Roy, Random function priors for exchangeable arrays with applications to graphs and relational data, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2012), pp. 998–1006.
 - [19] P. Orbanz and D. M. Roy, Bayesian models of graphs, arrays and other exchangeable random structures, *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 437 (2014).
 - [20] P. Hoff, Modeling homophily and stochastic equivalence in symmetric relational data, in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Red Hook, NY, 2008).
 - [21] D. J. Aldous, Representations for partially exchangeable arrays of random variables, *J. Multivar. Anal.* **11**, 581 (1981).
 - [22] D. N. Hoover, Relations on probability spaces and arrays of random variables, Technical Report (Institute for Advanced Study, Princeton, 1979), Vol. 2.

- [23] O. Kallenberg, Multivariate sampling and the estimation problem for exchangeable arrays, *J. Theor. Probab.* **12**, 859 (1999).
- [24] J. Besag, Spatial interaction and the statistical analysis of lattice systems, *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **36**, 192 (1974).
- [25] G. I. Allen and Z. Liu, A local poisson graphical model for inferring networks from sequencing data, *IEEE Trans. Nanobiosci.* **12**, 189 (2013).
- [26] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting, Poisson dependency networks: Gradient boosted models for multivariate count data, *Mach. Learn.* **100**, 477 (2015).
- [27] D. Strauss and M. Ikeda, Pseudolikelihood estimation for social networks, *J. Am. Stat. Assoc.* **85**, 204 (1990).
- [28] A. A. Amini, A. Chen, P. J. Bickel, E. Levina *et al.*, Pseudolikelihood methods for community detection in large sparse networks, *Ann. Stat.* **41**, 2097 (2013).
- [29] D. Garlaschelli and M. I. Loffredo, Patterns of Link Reciprocity in Directed Networks, *Phys. Rev. Lett.* **93**, 268701 (2004).
- [30] L. Peel, D. B. Larremore, and A. Clauset, The ground truth about metadata and community detection in networks, *Sci. Adv.* **3**, e1602548 (2017).
- [31] P. Gopalan, J. M. Hofman, and D. M. Blei, Scalable recommendation with hierarchical poisson factorization, in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Amsterdam, Netherlands, 2015), pp. 326–335.
- [32] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoidi, and A. Clauset, Stacking models for nearly optimal link prediction in complex networks, *Proc. Natl. Acad. Sci. USA* **117**, 23393 (2020).
- [33] C. De Bacco, D. B. Larremore, and C. Moore, A physical model for efficient ranking in networks, *Sci. Adv.* **4**, eaar8260 (2018).
- [34] Erasmus mobility statistics, <https://data.europa.eu/euodp/en/data>.
- [35] M. Contisciani, E. A. Power, and C. De Bacco, Community detection with node attributes in multilayer networks, *Sci. Rep.* **10**, 15736 (2020).
- [36] M. E. J. Newman and A. Clauset, Structure and inference in annotated networks, *Nat. Commun.* **7**, 11863 (2016).
- [37] T. Hoffmann, L. Peel, R. Lambiotte, and N. S. Jones, Community detection in networks without observing edges, *Sci. Adv.* **6**, eaav1478 (2020).
- [38] N. Stanley, T. Bonacci, R. Kwitt, M. Niethammer, and P. J. Mucha, Stochastic block models with multiple continuous attributes, *Appl. Network Sci.* **4**, 1 (2019).
- [39] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks, *Phys. Rev. X* **6**, 031005 (2016).
- [40] C. Blundell, J. Beck, and K. A. Heller, Modelling reciprocating relationships with Hawkes processes, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2012), Vol. 25.
- [41] X. Zhang, C. Moore, and M. E. Newman, Random graph models for dynamic networks, *Eur. Phys. J. B* **90**, 200 (2017).
- [42] S. Linderman and R. Adams, Discovering latent network structure in point process data, in *Proceedings of the 31st International Conference on Machine Learning*, edited by E. P. Xing and T. Jebara, Proceedings of Machine Learning Research (JMLR.org, Beijing, China, 2014), Vol. 32, pp. II-1413–II-1421.
- [43] T. P. Peixoto and M. Rosvall, Modelling sequences and temporal networks with dynamic community structures, *Nat. Commun.* **8**, 582 (2017).
- [44] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, *Science* **328**, 876 (2010).
- [45] P. Block, Reciprocity, transitivity, and the mysterious three-cycle, *Soc. Netw.* **40**, 163 (2015).
- [46] C. Seshadhri, A. Sharma, A. Stolman, and A. Goel, The impossibility of low-rank representations for triangle-rich complex networks, *Proc. Natl. Acad. Sci. USA* **117**, 5631 (2020).
- [47] T. P. Peixoto, Disentangling homophily, community structure and triadic closure in networks, [arXiv:2101.02510](https://arxiv.org/abs/2101.02510).
- [48] P. Block, C. Stadtfeld, and T. A. Snijders, Forms of dependence: Comparing saoms and ergms from basic principles, *Sociol. Meth. Res.* **48**, 202 (2019).
- [49] T. A. Snijders, Stochastic actor-oriented models for network change, *J. Math. Sociol.* **21**, 149 (1996).
- [50] T. A. Snijders, The statistical evaluation of social network dynamics, *Sociol. Methodol.* **31**, 361 (2001).
- [51] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, New specifications for exponential random graph models, *Sociol. Methodol.* **36**, 99 (2006).
- [52] <https://github.com/mcontisc/CRep>
- [53] Y. Zhu, X. Yan, L. Getoor, and C. Moore, Scalable text and link analysis with mixed-topic link models, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, NY, 2013), pp. 473–481.
- [54] J. Kunegis, KONECT: The Koblenz network collection, in *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion* (ACM Press, New York, NY, 2013), pp. 1343–1350.
- [55] H. Makse, POK Dataset, <https://hmake.cuny.cuny.edu/>.
- [56] A. Clauset, S. Arbesman, and D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks, *Sci. Adv.* **1**, e1400005 (2015).
- [57] J. Leskovec and A. Krevl, SNAP Datasets: Stanford large network dataset collection, <http://snap.stanford.edu/data> (2014).
- [58] P. Ji and J. Jin, Coauthorship and citation networks for statisticians, *Ann. Appl. Stat.* **10**, 1779 (2016).
- [59] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, Arnetminer: Extraction and mining of academic social networks, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08* (ACM Press, New York, NY, 2008), pp. 990–998.
- [60] V. A. Traag, L. Waltman, and N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities, *Sci. Rep.* **9**, 5233 (2019).

Community detection and reciprocity in networks by jointly modelling pairs of edges

MARTINA CONTISCIANI[†], HADISEH SAFDARI AND CATERINA DE BACCO

Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen 72076, Germany

[†]Corresponding author. Email: martina.contisciani@tuebingen.mpg.de

[Received on 23 February 2022; editorial decision on 28 June 2022; accepted on 20 July 2022]

To unravel the driving patterns of networks, the most popular models rely on community detection algorithms. However, these approaches are generally unable to reproduce the structural features of the network. Therefore, attempts are always made to develop models that incorporate these network properties beside the community structure. In this article, we present a probabilistic generative model and an efficient algorithm to both perform community detection and capture reciprocity in networks. Our approach jointly models pairs of edges with exact two-edge joint distributions. In addition, it provides closed-form analytical expressions for both marginal and conditional distributions. We validate our model on synthetic data in recovering communities, edge prediction tasks and generating synthetic networks that replicate the reciprocity values observed in real networks. We also highlight these findings on two real datasets that are relevant for social scientists and behavioural ecologists. Our method overcomes the limitations of both standard algorithms and recent models that incorporate reciprocity through a pseudo-likelihood approximation. The inference of the model parameters is implemented by the efficient and scalable expectation–maximization algorithm, as it exploits the sparsity of the dataset. We provide an open-source implementation of the code online.

Keywords: community detection; latent variables; network analysis; probabilistic generative models; reciprocity.

1. Introduction

Network models are powerful and flexible tools for representing complex interactions between individual elements in many different fields [1–4]. For instance, in social support networks, each individual is a person or the representative of a household, and each link, tie or arc represents the presence or intensity of a relationship between two individuals. Understanding what core patterns drive the observed set of interactions is of high relevance for scientists and practitioners willing to fully exploit the increased availability of networked datasets. A popular approach to modelling networks is that of generative models, in particular latent variable models [5]. They are probabilistic models that introduce latent variables to incorporate domain knowledge and capture complex interactions. Of particular interest, is the possibility of recovering clusters of individuals that behave similarly, a problem named community detection [6]. In this framework, the latent variables represent the nodes' community memberships and the structure of interactions between communities, and the aim is to infer these quantities from the data [7, 8]. Despite their flexibility and computational efficiency, these models have a main flaw: they fail in reproducing important structural network properties such as transitivity, reciprocity or triadic closure [9–11]. Synthetic networks generated from these models tend to have significantly lower values of these properties than those observed in real networks.

One possible reason of this problem is the common assumption of conditional independence: conditioned on the latent variables, network edges are independent and the joint probability distribution is

factorized accordingly. This means that an interaction from node i to node j is not directly affected by the interaction in the opposite direction, that is, the edge $j \rightarrow i$. In latent variable models with community structure, such as the stochastic block model [12] and its variants, these two edges are fully explained by the membership of the two nodes and sometimes by additional parameters such as degree corrections [13]. While this assumption has been used to obtain tractable problems, it can be too restrictive in certain real scenarios where non-trivial interaction patterns are observed. For instance, in social support networks, it is likely that the existence of interactions from individual i to individual j does not depend only on the groups that i and j belong to, but also on the fact that j has already previously helped i . This tendency of forming mutual connections is called reciprocity [14] and it is an important feature in social networks [15, 16], journal citations [17] and email communications [18, 19], to name a few. While exponential random graph models can represent such network properties in some form [20–23], they do not incorporate *a priori* latent variables as community membership. In the previous example, incorporating both community structure and the structural property of reciprocity would help us to understand how an individual interacts with others. Hence, there is a need to incorporate both these phenomena within a unique probabilistic framework.

Recently, Safdari *et al.* [10] tackled this problem by modelling the conditional distribution of *pairs* of edges between the same nodes, an assumption also shared by seminal works [12, 24]. Safdari *et al.* [10] include both communities and reciprocity effects inside the likelihood distribution of the network. This resulted in networks samples with values of reciprocity more similar to those of real data, and better edge predictions. However, this model relies on a pseudo-likelihood approximation for parameters' inference, as the model only specifies conditional distributions, but not the *joint* distribution of a pair of edges. As a result of this approximation, the model is not robust in community detection in the regime where reciprocity plays a role. Peixoto [9] has shown similar results in terms of triadic closure with a model based on Bayesian inference that combines community structure and this network property. This model also assumes conditional independence among edges and models conditional distributions of triadic edges.

Here, we propose a model that takes into account community structure and reciprocity by specifying a closed-form joint distribution of a pair of network edges, which does not involve approximations. To estimate the likelihood of network ties, we use a bivariate Bernoulli distribution—a special case of the multivariate Bernoulli distribution—where the log-odds are linked to community memberships, and pair-interaction variables. Although these patterns are indicative of two distinct mechanisms of network formation, namely, community structure and reciprocity, it is reasonable to expect that they are related to each other. For instance, (i) the preferred connection between nodes of the same community can induce the presence of reciprocated edges involving similar nodes and (ii) the tendency of forming mutual connections can induce the formation of groups of nodes. This conflation means that we cannot reliably interpret the underlying mechanisms of network formation merely from the abundance of reciprocated edges or observed community structure in network data. Our model takes advantage of the useful properties of the bivariate Bernoulli distribution, that is, the independence and the uncorrelatedness of the component random variables are equivalent and both the marginal and conditional distributions still follow the Bernoulli distribution. Hence, our model has closed-form analytical expressions and enables practitioners to address with more accuracy questions that were not fully captured by standard models; for instance, predicting the joint existence of mutual ties between pairs of nodes. In addition, its algorithmic implementation is efficient and scalable to large system size, as it exploits the sparsity of network datasets, thus allowing its broad applications across disciplines, for example, citation networks or neuronal networks that consist of several thousand of nodes.

2. The model

The main goal of this work is to develop a probabilistic generative model with latent variables that better captures real scenarios where non-trivial interaction patterns are observed in networks. This is achieved by modelling *jointly* the edges between the same pair of nodes, differently from standard models that assume their conditional independence given the latent variables. Formally, we model the interactions of N individuals as a binary asymmetric matrix A , with entries A_{ij} defining the presence or the absence of connections from node i to node j . Our model considers jointly the pair $A_{(ij)} := (A_{ij}, A_{ji})$ distributed with a bivariate Bernoulli distribution of parameters Θ , which takes values from $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ in the Cartesian product space $\{0, 1\}^2 = \{0, 1\} \times \{0, 1\}$. Its probability density function can be written as

$$\begin{aligned} P(A_{(ij)}|\Theta) &= P(A_{ij}, A_{ji}|\Theta) \\ &= p_{11}^{A_{ij}A_{ji}} p_{10}^{A_{ij}(1-A_{ji})} p_{01}^{(1-A_{ij})A_{ji}} p_{00}^{(1-A_{ij})(1-A_{ji})} \\ &= \frac{\exp\{A_{ij}f_{ij} + A_{ji}f_{ji} + A_{ij}A_{ji}J_{(ij)}\}}{Z_{(ij)}}, \end{aligned} \quad (2.1)$$

where $Z_{(ij)}$ is a normalization constant and $p_{00} = 1/Z_{(ij)}$. In addition, $p_{00} + p_{10} + p_{01} + p_{11} = 1$, and

$$f_{ij} = \log\left(\frac{p_{10}}{p_{00}}\right), f_{ji} = \log\left(\frac{p_{01}}{p_{00}}\right), J_{(ij)} = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right). \quad (2.2)$$

Thus, $P(A_{ij}, A_{ji}|\Theta)$ can be viewed as a member of the exponential family, and can be represented in a log-linear formulation as in equation (2.1), where f_{ij}, f_{ji} , and $J_{(ij)}$ represent the natural parameters. $J_{(ij)}$ is called cross-product ratio between A_{ij} and A_{ji} and represents the log-odds of the model. Similar to the Ising model [25], if $J_{(ij)} = 0$ then the components of the bivariate Bernoulli random vector (A_{ij}, A_{ji}) are independent, thanks to the equivalence of independence and uncorrelatedness for multivariate Bernoulli distributions [26]. In this case, the resulting model would be equivalent to consider the product of two independent Bernoulli distributions. Another interesting property of the bivariate Bernoulli is that both marginal and conditional distributions are univariate Bernoulli. Thus, our model has closed-form equations for joint, conditional and marginal distributions.

We now assume that a set of latent variables capture hidden patterns of the data. There are many possibilities to add these variables: one could act directly on the marginal or conditional first moments, as well as modelling separately the different $p_{\alpha\beta}$, with $\alpha, \beta \in \{0, 1\}$. However, we model the log-ratios to ease interpretability and the analytical computations. Specifically, we assume

$$f_{ij} = \log \lambda_{ij} \quad (2.3)$$

$$f_{ji} = \log \lambda_{ji} \quad (2.4)$$

$$J_{(ij)} = \log \eta, \quad (2.5)$$

where

$$\lambda_{ij} = \sum_{k,q=1}^K u_{ik} v_{jq} w_{kq} \quad (2.6)$$

captures mixed-membership community structure as in De Bacco *et al.* [8] and η is the pair-interaction coefficient. The parameters u_{ik}, v_{jq} are entries of K -dimensional vectors \mathbf{u}_i and \mathbf{v}_i , the out-going and incoming communities, respectively; and w_{kq} are the entries of a $K \times K$ affinity matrix, which regulates the structure of communities, for example, assortative when its diagonal entries are greater than off-diagonal entries (homophily). Thus, $\Theta = (u, v, w, \eta)$ are the latent parameters we want to infer. Through equations (2.3)–(2.5), we encode the assumptions that community structure drives the process of edge formation, and the edges of a pair of nodes depend on each other explicitly according to the parameter η . When $J_{(ij)} = 0$, the probability of $A_{(ij)}$ is given by the agreements of the communities of i and j only; while a positive value for the log-odds will boost the chance to observe a tie between them. Conversely, $J_{(ij)} < 0$ decreases the value of p_{11} , the probability that both edges exist. Considering equation (2.5), $0 < \eta < 1$ and $\eta > 1$ codify a negative and positive interaction between i and j , respectively. The first lowers the probability of observing both ties $i \rightarrow j$ and $j \rightarrow i$, while the latter increases it. Finally, $\eta = 1$ implies no interaction between A_{ij} and A_{ji} . With this model at hand we can estimate observable quantities, valuable for practitioners. For instance, one can ask about the expected value of a given tie in general or conditioned on the existence of the opposite one, quantities defined as:

$$\mathbb{E}[A_{ij}] = \frac{\lambda_{ij} + \eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}}, \quad (2.7)$$

$$\mathbb{E}[A_{ij}|A_{ji}] = \frac{\eta^{A_{ji}} \lambda_{ij}}{\eta^{A_{ji}} \lambda_{ij} + 1}, \quad (2.8)$$

and similar for $\mathbb{E}[A_{ji}]$ and $\mathbb{E}[A_{ji}|A_{ij}]$, see Appendix A. With these quantities one can perform edge prediction tasks, which is crucial when we are limited to a subset of the dataset.

3. Inference

We infer the parameters using a maximum likelihood approach. Specifically, we maximize the log-likelihood

$$\mathcal{L}(\Theta) = \sum_{ij} f_{ij} A_{ij} + \frac{1}{2} \sum_{ij} J_{(ij)} A_{ij} A_{ji} - \frac{1}{2} \sum_{ij} \log Z_{(ij)} \quad (3.1)$$

with respect to $\Theta = (u, v, w, \eta)$. Adopting a variational approach, this is equivalent to maximize

$$\begin{aligned} \mathcal{L}(\rho, \Theta) = & \sum_{ij} \left[A_{ij} \sum_{k,q} \rho_{ijkq} \log \left(\frac{u_{ik} v_{jq} w_{kq}}{\rho_{ijkq}} \right) + \frac{1}{2} A_{ij} A_{ji} \log \eta \right. \\ & \left. - \frac{1}{2} \log \left(\sum_{k,q} u_{ik} v_{jq} w_{kq} + \sum_{k,q} u_{jk} v_{iq} w_{kq} + \eta \sum_{k,q} u_{ik} v_{jq} w_{kq} \sum_{k,q} u_{jk} v_{iq} w_{kq} + 1 \right) \right], \end{aligned} \quad (3.2)$$

where we introduced the variational distribution ρ_{ijkq} over the parameters and used Jensen's inequality. The equivalence holds when

$$\rho_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k,q} u_{ik} v_{jq} w_{kq}}. \quad (3.3)$$

Algorithm 1 JointCRep: EM algorithm**Input:** network $A = \{A_{ij}\}_{i,j=1}^N$, number of communities K .**Output:** membership matrices $u = [u_{ik}]$, $v = [v_{ik}]$; network-affinity matrix $w = [w_{kq}]$; pair-interaction parameter η .Initialize u, v, w, η at random.Repeat until \mathcal{L} convergences:

1. Calculate ρ (E-step):

$$\rho_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k,q} u_{ik} v_{jq} w_{kq}}$$

2. Update parameters Θ (M-step):

- (i) for each pair (i, k) update memberships:

$$u_{ik} = \frac{\sum_{j,q} A_{ij} \rho_{ijkq}}{\sum_j \left[\frac{\sum_q v_{jq} w_{kq} (1 + \eta \lambda_{ji})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}$$

$$v_{ik} = \frac{\sum_{j,q} A_{ji} \rho_{jiqk}}{\sum_j \left[\frac{\sum_q u_{jq} w_{qk} (1 + \eta \lambda_{ij})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}$$

- (ii) for each pair (k, q) update affinity matrix:

$$w_{kq} = \frac{\sum_{i,j} A_{ij} \rho_{ijkq}}{\sum_{i,j} \left[\frac{u_{ik} v_{jq} (1 + \eta \lambda_{ji})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}$$

- (iii) update pair-interaction parameter:

$$\eta = \frac{\sum_{i,j} A_{ij} A_{ji}}{\sum_{i,j} \left[\frac{\lambda_{ij} \lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}$$

We estimate the parameters by using an expectation–maximization (EM) algorithm where at each step one updates ρ using equation (3.3) (E-step) and then maximizes $\mathcal{L}(\rho, \Theta)$ with respect to $\Theta = (u, v, w, \eta)$ by setting partial derivatives to zero (M-step). This iteration is repeated until the log-likelihood converges. The exact equations for the updates of the parameters are in Appendix A, and the whole routine is described in Algorithm 1. This algorithm is computationally efficient and scalable to large system sizes as it exploits the sparsity of the dataset. Indeed, all the updates involved in the numerator sum over A_{ij} , hence only the non-zero entries count, giving an algorithmic complexity of $O(M K^2)$, where $M = \sum_{i,j} A_{ij}$ is the number of ties.

Our model (JointCRep) aims to generalize the method presented in Safdari *et al.* [10] (CRep), which was of inspiration for the latent variables underlying the generative process. We refer to [10] for a detailed explanation of this method and summarize its main properties in Table 1.

TABLE 1 *Properties of JointCRep, CRep and MT models. λ represents the community effect and η is the parameter linked to the reciprocity \mathfrak{r} . MT is a community detection-only model, therefore it does not have a reciprocity parameter. In addition, it uses the conditional independence assumption according to which the conditional and the marginal distributions coincide. For this reason, the closed-form conditional and joint do not apply for this method*

	JointCRep	CRep	MT
Networks	Binary	Weighted	Weighted
Likelihood	Bivariate Bernoulli	Poisson	Poisson
Marginal mean	$\mathbb{E}[A_{ij}] = \frac{\lambda_{ij} + \eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}}$	$\mathbb{E}[A_{ij}] = \frac{\lambda_{ij} + \eta \lambda_{ji}}{1 - \eta^2}$	$\mathbb{E}[A_{ij}] = \lambda_{ij}$
Conditional mean	$\mathbb{E}[A_{ij} A_{ji}] = \frac{\eta^{A_{ji}} \lambda_{ij}}{\eta^{A_{ji}} \lambda_{ij} + 1}$	$\mathbb{E}[A_{ij} A_{ji}] = \lambda_{ij} + \eta A_{ji}$	$\mathbb{E}[A_{ij} A_{ji}] = \mathbb{E}[A_{ij}]$
Relationship η vs. \mathfrak{r}	Sublinear	Linear	–
Contribution λ vs. η	Multiplicative	Additive	–
Contribution \mathfrak{r}	Real	Non-negative	–
Closed-form marginal	Yes	No	Yes
Closed-form conditional	Yes	Yes	–
Closed-form joint	Yes	No	–

4. Results

In this section, we present the results obtained in synthetic and real networks. For comparison we use CRep, the model that combines communities and reciprocity with a pseudo-likelihood approximation [10], and MULTITENSOR (MT), a community detection-only generative model with a maximum likelihood approach [8]. Even if both of them posit a Poisson likelihood, in this work, we use only binary networks for fair comparisons with our model JointCRep. We summarize the main similarities and differences among the models used in the analysis in Table 1.

4.1 Results on synthetic data

We first validate the performance of the different methods on synthetic data generated with the model proposed in this work. Being a generative model, given as input an initial set of parameters, one can draw a directed network with a community structure and a reciprocity value from the expression in Equation (2.1). The generative process is described in detail in Appendix B. We analyse networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of the pair-interaction parameter η such that we obtain networks with reciprocity values (\mathfrak{r}) in the interval $[0, 0.8]$. We generate 10 random samples for each value of \mathfrak{r} . In addition to these results, we provide further details for synthetic networks generated with different values of average degree in Appendix C.3. We test the ability of the models to (i) recover the communities, (ii) perform edge prediction tasks and (iii) generate sample networks that replicate relevant network quantities.

4.1.1 Community detection To evaluate the performance of the methods on recovering the communities, we use the cosine similarity (CS), a measure useful to capture mixed-membership communities, as in this case. It ranges from 0 to 1, where 1 means perfect recovery. We calculate the average of the cosine

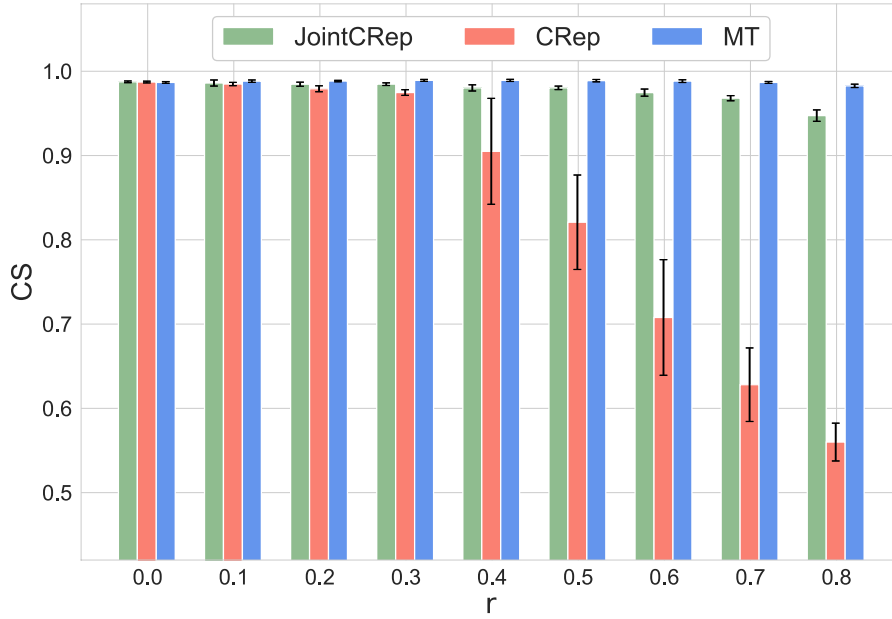


FIG. 1. Community detection in synthetic networks. Cosine similarity (CS) in synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of reciprocity r . Results are averages and standard deviations over 10 synthetic networks.

similarities of both membership matrices u and v , and then averaging over the nodes. The results are shown in Fig. 1. In the scenarios with low reciprocity values ($r < 0.4$), all models perform well. However, as r increases, CRep worsens while JointCRep keeps having good results comparable to those of the community-only algorithm, MT. The big drop of CRep is due to the fact that this model gives increasingly less weight to communities as reciprocity increases, as pointed out in Safdari *et al.* [10]. Conversely, JointCRep is not affected by the different reciprocity values of the data and still performs as good as MT, even by adding another parameter to the model.

4.1.2 Edge prediction The edge prediction task consists in estimating the existence of an edge by using the inferred parameters. The main quantity used as a score for the estimation of the entries of the adjacency matrix A is the expected value of the marginal distribution. However, our model also provides the conditional distribution; hence, its expected value can also be used as a score. The difference lies in the nature of the question we try to answer. We use the marginal distribution to merely predict the existence of an edge, without considering additional information. On the other hand, with the conditional distribution, we ask what is the probability of an edge $i \rightarrow j$, conditioned on observing the state of the opposite edge $j \rightarrow i$, that is, knowing if it exists or not. Here, we exploit the presence or the absence of the edge in the opposite direction to better predict each given entry. Furthermore, our model specifies a joint distribution over the edges of a pair of nodes, and this allows us to answer questions more accurately compared to the standard models, which do not specify a joint distribution. For instance, what is the probability of *jointly* observing both edges or even only an edge in one direction while not observing the other in the opposite? Our model directly captures this by specifying $P(A_{ij}, A_{ji} | \Theta)$, while others positing a conditional independence assumption can only compute an approximation as $P(A_{ij} | \Theta) P(A_{ji} | \Theta)$.

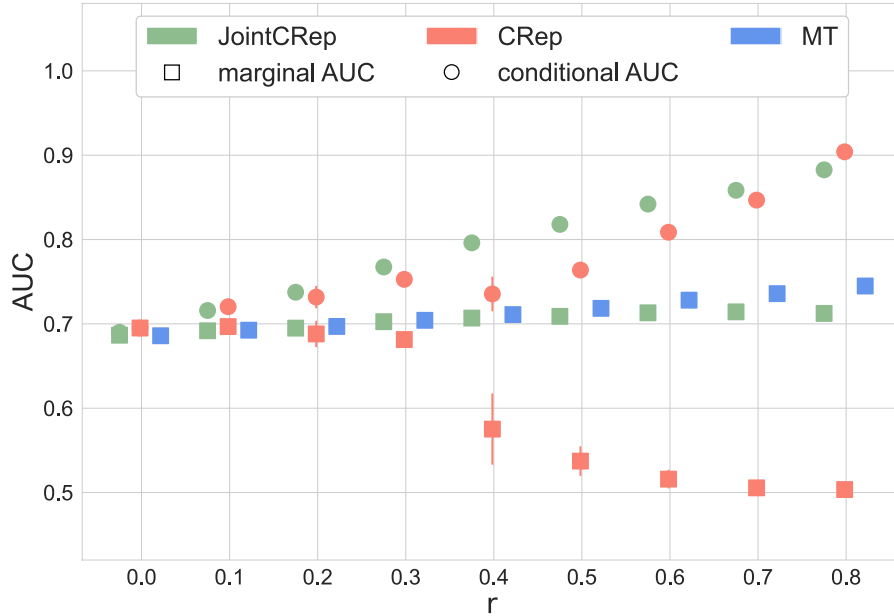


FIG. 2. Edge prediction in synthetic networks. Synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of reciprocity r . Results are averages and standard deviations over 10 synthetic networks and over 5-folds of cross-validation test sets. Edge prediction performance is measured with AUC and the baseline is the random value 0.5.

In our experiments below, we test edge prediction with various scores by using 5-fold cross-validation. Specifically, we divide the dataset into five equal-size groups (folds) and train the models on four of them (training data) for learning the parameters; this contains 80% of the possible pairs of nodes in the network, so that we hide pairs of entries (A_{ij}, A_{ji}) from the training. One then predicts the existence of edges in the held-out group (test set). As performance metric, we measure the AUC on the test data, that is, the probability that a randomly selected edge has higher expected value than a randomly selected non-existing edge. We compute both the regular, and conditional AUC values. To estimate the regular AUC, we take the expected value $\mathbb{E}_{P(A_{ij}|\ominus)}[A_{ij}]$ as the score; while for the conditional AUC, the expected value over the conditional distribution, that is, $\mathbb{E}_{P(A_{ij}|A_{ji},\ominus)}[A_{ij}]$ acts as the score. The latter cannot be computed for the community detection-only algorithm, as the marginal distribution is the same as the conditional, and thus the two AUC values coincide. We provide more details in Appendix C.1, where we also show the ability of our model on edge prediction tasks by using the joint distribution.

Figure 2 displays the results of the marginal and conditional edge prediction for the different models. JointCRep significantly improves the performance of CRep when using the marginal expected value, and it performs as good as MT. The latter, however, is not able to exploit the additional information given by the existence (or non-existence) of the edge in the opposite direction. This dependence is crucial in networks with reciprocity, that is, most of the real world datasets, and models with an explicit conditional distribution can better adopt this information to obtain higher performance in edge prediction. Indeed, JointCRep and CRep perform remarkably in this task, and our model presents more robust results both in terms of standard deviation and growth.

4.1.3 Reproducing the topological properties A notable property of generative models is their ability to produce synthetic networks based on real-world datasets, such that the synthetic networks imitate

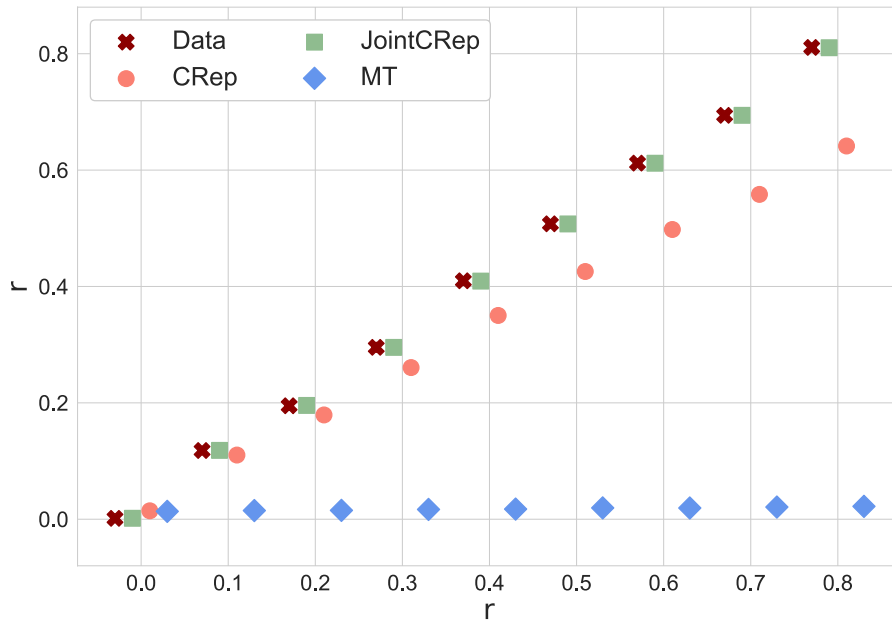


FIG. 3. Reciprocity in sampled synthetic networks. Synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of reciprocity r . Results are empirical averages and standard deviations over 50 samples of 10 independent synthetic networks (five samples per input network). We measure the reciprocity and the cross (x) markers indicate the average on 10 input networks.

the topological features of the real datasets. Following the approach in Safdari *et al.* [10], for each individual network, we infer the network parameters by applying each model. Then, we use these inferred parameters to generate five network samples. We compare topological properties of these samples with those observed in the ground truth networks used to infer the parameters. In particular, we are interested in measuring reciprocity. Figure 3 shows the performance of each model in reproducing this feature in sampled networks. As it is expected, MT is not capable of reflecting the observed value of the reciprocity in the ground truth network, a clear indication of the shortcoming of models based purely on community structure, which indeed limits their applications. Conversely, JointCRep perfectly reproduces this quantity. CRep generates sampled networks with reciprocity lower than the ground truth due to the fact that it uses a Poisson likelihood resulting in weighted networks. Additional results are provided in Appendix C.2.

To summarize the results on synthetic networks, JointCRep is capable of recovering communities on networks with varying reciprocity values, performing as good as models that are based purely on community structure. This capability overcomes the limitations of the recent CRep model. Moreover, JointCRep includes many performance enhancements in the edge prediction task, that is, showing improved results in terms of marginal AUC and more robust conditional AUC values. Furthermore, JointCRep is also capable of generating sampled networks with topological features that resemble that of the real data, for example, reciprocity and average degree. Collectively, these findings show that JointCRep is able to overcome the limitations of both the community detection-only algorithm MT and the model that incorporates reciprocity through the pseudo-likelihood approximation CRep.

4.2 Analysis of a high-school social network

We now study the social network that describes friendships between boys in a small high-school in Illinois that was collected in the fall of 1957 [27]. Here, a node represents a boy and an edge from an individual i

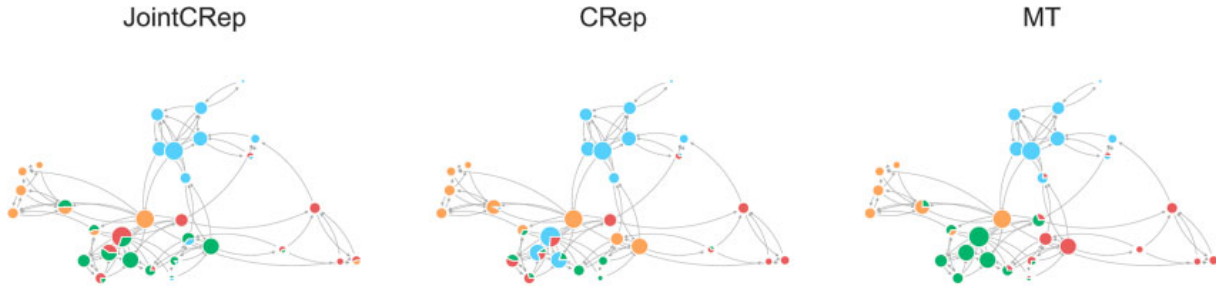


FIG. 4. Community detection in the high-school social network. Mixed-membership partitions determined by the matrix u inferred by JointCRep, CRep and MT. Node size is proportional to the degree (in- and out-degree).

to j shows that node i claimed to be friend of node j . We pre-processed the dataset by removing self-loops and isolated nodes. The resulting directed network has 31 nodes, 100 edges and reciprocity equal to 0.52, that is, only half of the edges (friendship relationships) are reciprocated. There is no additional metadata to describe the nodes, nor is there an available ground truth for the latent parameters. Therefore, we estimate the number of communities K by performing edge prediction tasks via 5-fold cross-validation with different values of K . For each method the best performance in terms of AUC was achieved with $K = 4$. Edge prediction also serves as model validation routine in the absence of ground truth information, as it is the case here. We found that results vary depending on the metric considered for evaluation, but in general they confirm that all models are fitting the data well, considering that the dataset is small and highly sparse, thus making prediction tasks hard. Further details for the edge prediction task are in Appendix D.2. Figure 4 visualizes the mixed-membership partitions resulting from the matrix u , inferred by the different methods (similar results are obtained for v). Here we use the inferred value of u , which is obtained from the run with the highest log-likelihood over 100 random initializations of the parameters. All the algorithms assign most of the students to the same groups, except from a central block. Here, MT infers mostly hard memberships and balances the number of nodes in each cluster. Instead, CRep allocates only three nodes with small degree to the green community while places the nodes with higher degree in other clusters. JointCRep, shows a partition that lies in between, by inferring mixed-memberships for those nodes known as *bridges*. To measure quantitatively the diversity of communities inferred by the various methods, we compute a modularity for directed networks and overlapping communities using different aggregation functions, as proposed by Nicosia *et al.* [28]. Modularity assumes all communities to have statistically similar properties, in particular to have similar sizes, and it is suited for assortative community structure [29, 30]. The communities shown in Fig. 4 reflect these properties to a certain extent, and we report the results in terms of overlapping modularity in Table D2. We find modularity values between 0.48 and 0.75, depending on the aggregation function, with JointCRep and MT achieving the highest values.

Given the inferred parameters, we can test the ability of the models to reconstruct the input network, by using either the marginal expected value $\mathbb{E}_{P(A_{ij}|\Theta)}[A_{ij}]$, or the conditional expected value $\mathbb{E}_{P(A_{ij}|A_{ij},\Theta)}[A_{ij}]$ as the score. Note that the latter is not available for MT because the conditional and marginal distributions coincide. Figure 5 presents the results, where edge width and darkness of the reconstructed networks are proportional to the weight given by the expected score (for visualization clarity, we show only edges with weight greater than 0.2). The network estimated by CRep, which uses the expected value of the marginals, does not capture the structure of the data magnificently, as it overestimates the presence of edges. This model specifies conditional distributions and relies on a pseudo-likelihood approximation; since this approach is not necessarily accurate enough to approximate marginals, such results are expected. Instead,

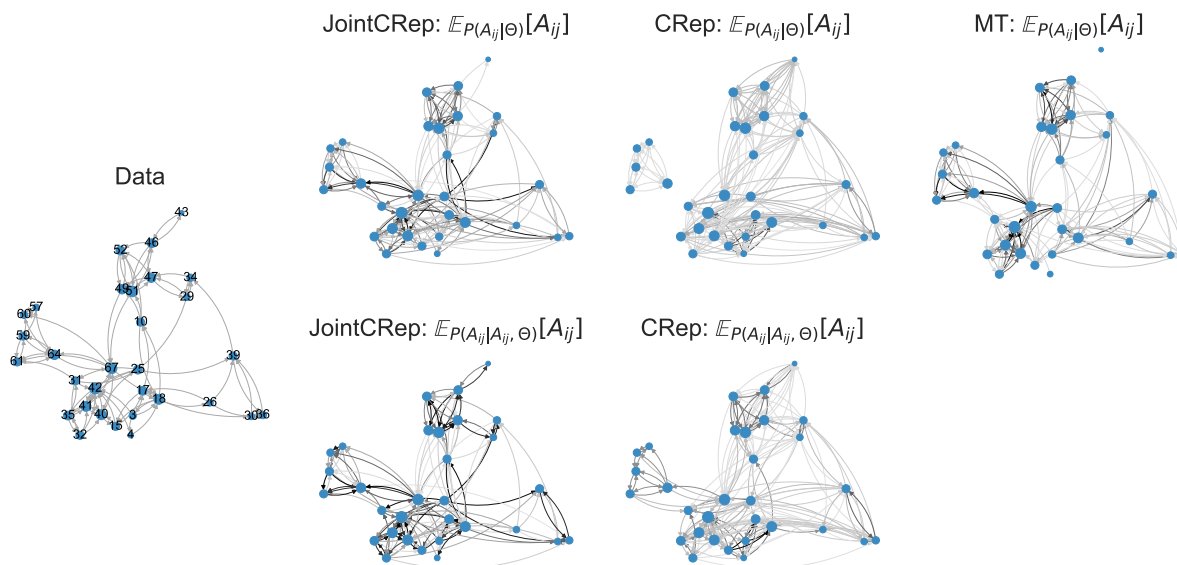


FIG. 5. High-school network reconstruction. (Left) High-school data and (right) network reconstructions by using as a score either the marginal expected value $\mathbb{E}_{P(A_{ij}|\Theta)}[A_{ij}]$ or the conditional expected value $\mathbb{E}_{P(A_{ij}|A_{ij}, \Theta)}[A_{ij}]$ with the inferred parameters. Note that the last is not available for MT because the conditional and marginal distributions coincide. Edge width and darkness are proportional to the weight (given by the expected score); for visualization clarity, we show only edges with weight greater than 0.2. Node size is proportional to the degree (in- and out-degree) and node labels represent node IDs.

MT and JointCRep estimate a sparser representation that is closer to the observed network. However, MT is not able to notably detect reciprocated edges, for example, (10, 18) or (64, 67), while JointCRep remarkably recovers this type of interactions more precisely. For both JointCRep and CRep, including the conditional expected values improves their accuracy in reconstruction, resulting in identifying reciprocal edges correctly. The difference between the two models lies on the intensity: for instance JointCRep predicts the pair of edges between nodes 10 and 18 with a high probability, while CRep assigns a much lower probability to them. Hence, JointCRep is not only able to predict edges more precisely, but it also does so with higher probability. These qualitative observations are also confirmed by quantitative comparisons in terms of the Log Loss and the L1 Loss, two penalty metrics computed between the reconstructed and the true networks. They measure the difference between two input networks by taking into account the probability of the existence of an edge and computing a penalty for each mistake in predicting the observed value. A penalty of 0 denotes perfect reconstruction, as when assigning probability 1 to the observed edge values. In general, lower values indicate higher similarity. Further details can be found in Appendix D.1. We find that JointCRep achieves the lowest values (i.e. better performance) in both metrics, with best overall performance achieved using the conditional expectation. See Table D3 for details.

To further compare the strength of these methods, we examine their performance in generating samples that resemble the observed network. For each model, we use the inferred parameters to generate five synthetic networks, as shown in Fig. 6. Again, we notice how the samples generated by JointCRep better resemble the observed network, as it is easier to distinguish the four blocks generated by JointCRep, compared to the samples from the other algorithms. In particular, JointCRep finds denser groups given by reciprocated edges. In addition to these qualitative results, Table D4 reports the topological properties of the observed data and the sampled networks, showing that JointCRep generates networks samples that on average are most similar to the observed data in terms of average degree, reciprocity and clustering coefficient.

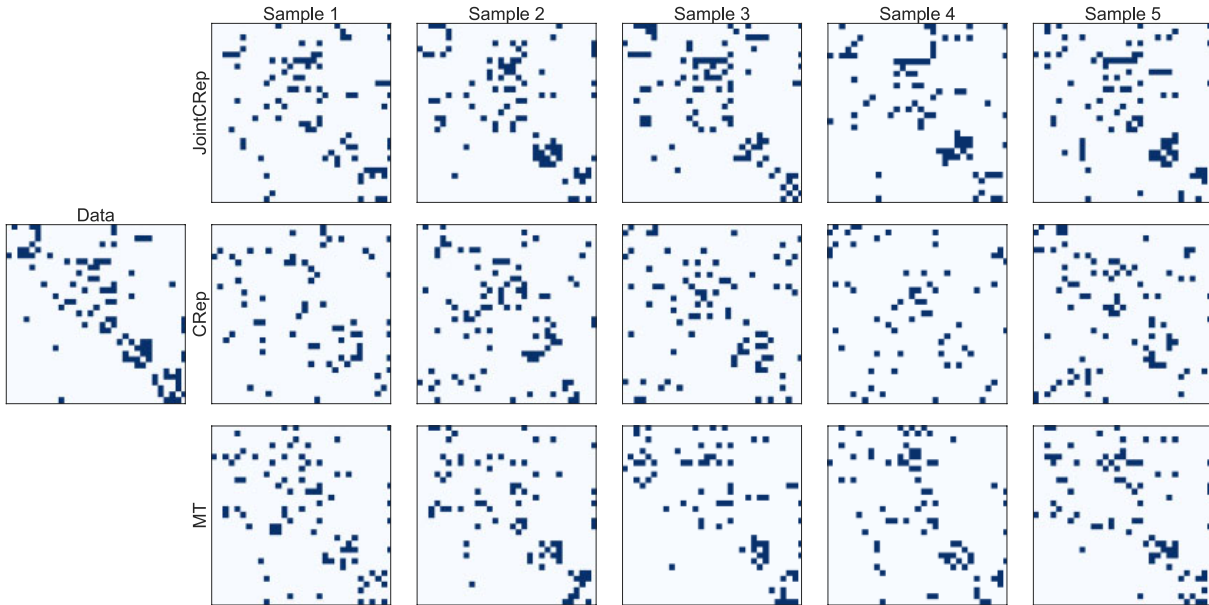


FIG. 6. High-school network samples. (Left) High-school data and (right) five random samples generated by different methods with the inferred parameters. For each method, we first infer the parameters choosing the run with the highest log-likelihood over 100 random initializations of the parameters. Then, we generate the samples by using as input in the generative models the inferred parameters and the average degree of the original data. The generative process of JointCRep is described in Appendix B; for CRep we use the formulation described in [10]; MT follows the formulation of a standard mixed-membership variant of a stochastic block model, as described in [8].

4.3 Analysis of a vampire bat network

As a second example, we study the network of food sharing interactions in captive vampire bats, collected by Carter and Wilkinson [31]. These animals often regurgitate food to roost-mates that fail to feed. The decision of who to feed may depend on both kin relatedness and reciprocal sharing. Hence, in this dataset, we expect reciprocity to be an important factor for tie formation. In the study, they fasted 20 vampire bats and induced food sharing on 48 days, over a 2-year period. They showed that reciprocal sharing predicts future food regurgitation more than relatedness or other non-kin food sharing behaviours, such as harassment. From the collected data, we construct a directed network by building an edge from a bat i to another j if node i fed j at least once. We removed isolated nodes and obtained a network with 19 nodes, 103 edges and reciprocity equal to 0.64. We fix the number of communities $K = 2$ and analyse the data with the different methods. As for the high-school data, results of edge prediction tasks for model validation confirm that all the models represent the data well, see Table D5 for details. We are now interested in measuring the ability of the models to recover the observed network with the inferred parameters, in particular their ability to recover topological properties such as reciprocity. To this aim, we consider the marginal and the conditional expected values, as in Section 4.2. Figure 7 shows the adjacency matrix of the data and its different estimates, obtained by each method. The network embodies a core-periphery structure, where the main core (labels 0–9) is made of female bats. JointCRep recovers this structure more accurately than other methods, the off-diagonal entries show this fascinating result clearly, while the other methods overestimate the amount of edges. Similarly as observed in the high-school network, our model is not only more accurate, but also assigns higher probabilities to these entries and best performs both in terms of Log and L1 Loss, see Table D6.

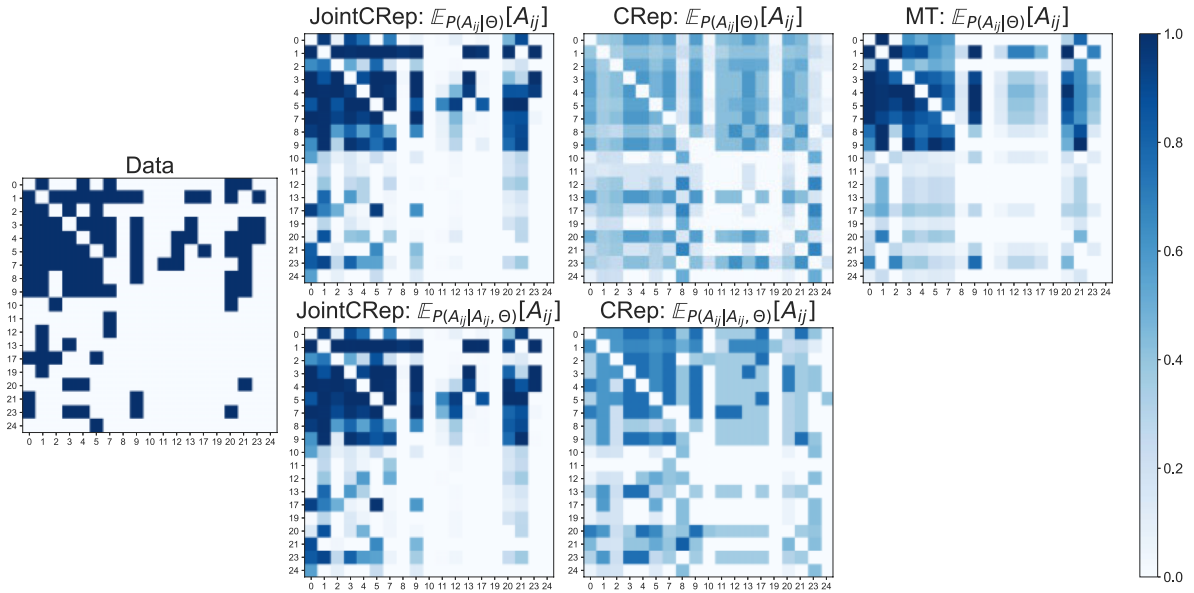


FIG. 7. Vampire bat network reconstruction. (Left) The adjacency matrix of the vampire bat data and (right) its estimates by using as a score either the marginal expected value $\mathbb{E}_{P(A_{ij}|\Theta)}[A_{ij}]$ or the conditional expected value $\mathbb{E}_{P(A_{ij}|A_{ij}, \Theta)}[A_{ij}]$ with the inferred parameters. Note that the last is not available for MT because the conditional and marginal distributions coincide. The intensity of the entries is proportional to the score probability, as shown in the colourbar. The labels near the ticks represent node IDs.

In addition to the marginal and conditional expected value, we can consider the joint distribution to estimate the entries of the adjacency matrix. This is equivalent to assign a value to each pair (A_{ij}, A_{ji}) from the set $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$, that transforms the edge prediction task into a classification problem. We predict the label associated to the highest probability among $[p_{00}, p_{01}, p_{10}, p_{11}]$, where these are computed by using equations (A.1)–(A.4) with the inferred parameters. We assess the goodness of our performance by computing the precision and recall of the predicted labels versus the true labels, as shown in Fig. 8. The precision identifies the proportion of correctly classified observed entries. The figure illustrates high precision values consistently across edge labels, as the highest entries are along the diagonal. In particular, JointCRep is able to correctly classify the pairs $(0, 0)$ and $(1, 1)$. Observing where our model misclassifies, this mainly happens by predicting no edges, that is, assign label $(0, 0)$, when the true ones are either $(0, 1)$ or $(1, 0)$, implying a tendency to estimate sparser networks. On the other hand, the recall indicates the proportion of predicted edges being correctly classified. Also in this case, the highest entries are in the main diagonal and in predicting the pairs $(0, 0)$ and $(1, 1)$. Overall, in this case, we obtain higher intensities as for the precision, indicating the tendency of labelling the predicted edges with the right type.

To conclude our analysis, we compare five random samples generated with the inferred parameters of each model and check whether they reproduce topological properties as those observed in the real data. Table 2 shows that JointCRep outperforms other models in terms of all topological properties. In particular, it generates sampled networks with reciprocity values closest to the real data but also reproduces realistic values of the clustering coefficient.

5. Discussion and conclusion

In this article, we have presented a generative model called JointCRep that takes into account community structure and reciprocity by specifying a closed-form joint distribution of a pair of network edges, without

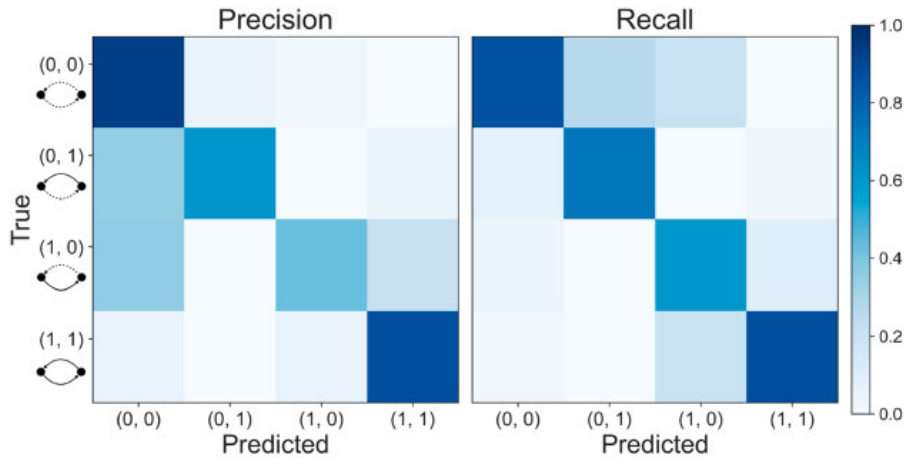


FIG. 8. Precision and recall of the vampire bat network. Statistics based on the confusion matrix that compares the entries of the adjacency matrix and the estimates obtained with the joint distribution of JointCRep. The precision is given by a normalization by row, while the recall accounts for the normalization by column. The label (0, 0) denotes no interactions between nodes i and j ; labels (0, 1) and (1, 0) considers the pair of edges where only one edge in one direction is present, and the label (1, 1) indicates reciprocated edges.

TABLE 2 *Topological properties in vampire bat and its sampled networks. Results are averages and standard deviations over five samples. We measure the number of nodes N , the number of edges M , the average degree $\langle k \rangle$, the reciprocity r and the clustering coefficient cc*

	N	M	$\langle k \rangle$	r	cc
Data	19	103	10.84	0.64	0.54
JointCRep	18.4 ± 0.89	100.4 ± 5.41	10.92 ± 0.38	0.61 ± 0.03	0.55 ± 0.05
CRep	18.2 ± 0.84	74.2 ± 5.40	8.16 ± 0.54	0.51 ± 0.04	0.27 ± 0.05
MT	17.4 ± 1.14	70.0 ± 7.38	8.06 ± 0.83	0.36 ± 0.06	0.37 ± 0.01

relying upon approximations. Our model also provides closed-form analytical expressions for both the marginal and conditional distributions, and enables practitioners to address with more accuracy questions that were not fully captured by standard models; for instance, predicting the joint existence of mutual ties between pairs of nodes.

We first validated our model by applying it to synthetic network datasets, where we achieved remarkable performance in recovering communities, edge prediction tasks and generating synthetic networks that replicate topological features observed in real networks. We then analysed two real datasets that are relevant for social scientists and behavioural ecologists, where we found that JointCRep obtains more robust and interpretable results. The results shown in this work highlight main benefits of using a model that considers closed-form joint distributions of pairs of edges in networks, while also showing possible shortcomings of other approaches. While it is difficult to pinpoint theoretical reasons for these shortcomings, the variety of experiments that we discussed throughout this manuscript show possible practical consequences of them. In particular, standard generative models make strong conditional independence assumptions that reflect in poor recovery of topological properties as reciprocity. On the other hand,

models that specify only conditional distributions rely on pseudo-likelihood approximations that may reflect in weak recovery of latent parameters as communities in certain regimes. Collectively, our model is able to overcome the limitations of both these approaches thanks to the modelling of closed-form joint distributions while also keeping computational complexity under control.

The framework we described could be extended in a number of ways. JointCRep analyses binary and single-layer networks; therefore, possible extensions could account for weighted and possibly multilayer networks, where we have edges of different types. Another approach could consider dynamic networks, which have evolving structure over time, and adapt the parameters accordingly [32]. Moreover, our model captures the reciprocity through a unique pair-interaction parameter for the whole network. This model could be improved in the future by including node-dependent parameters in scenarios where reciprocity varies between individuals. Furthermore, many real-world datasets contain attributes that provide additional information about their features. Incorporating these extra informations on nodes could result in a more realistic analysis [33].

JointCRep, to the best of our knowledge, is the first such method for fully capturing reciprocity by jointly modelling pairs of edges with exact two-edge joint distributions. We believe it will serve as a baseline for future models that tackle more complicated interactions that go beyond pairwise interaction, for example, triadic closure [9].

Funding

All the authors were supported by the Cyber Valley Research Fund.

Acknowledgements

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani.

Data and Code

An open-source algorithmic implementation of the model together with the code to generate synthetic data is publicly available and can be found at <https://github.com/mcontisc/JointCRep>.

Appendix

A. Detailed derivations

Combining equations (2.2)–(2.5), we get the explicit mapping between the latent variables and the instances of the joint probability in equation (2.1):

$$p_{01} = \frac{\lambda_{ji}}{Z_{(ij)}} \quad (\text{A.1})$$

$$p_{10} = \frac{\lambda_{ij}}{Z_{(ij)}} \quad (\text{A.2})$$

$$p_{11} = \frac{\eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}} \quad (\text{A.3})$$

$$p_{00} = \frac{1}{Z_{(ij)}}, \quad (\text{A.4})$$

where the normalization constant is:

$$Z_{(ij)} = \lambda_{ij} + \lambda_{ji} + \eta\lambda_{ij}\lambda_{ji} + 1. \quad (\text{A.5})$$

One property of the bivariate Bernoulli is that both marginal and conditional distributions are univariate Bernoulli. Thus, the marginal distributions of A_{ij} and A_{ji} have densities:

$$P(A_{ij}) = (p_{10} + p_{11})^{A_{ij}} (p_{00} + p_{01})^{(1-A_{ij})} \quad (\text{A.6})$$

$$P(A_{ji}) = (p_{01} + p_{11})^{A_{ji}} (p_{00} + p_{10})^{(1-A_{ji})}, \quad (\text{A.7})$$

while the conditional distributions are the following:

$$P(A_{ij}|A_{ji}) = \left(\frac{p(1, A_{ji})}{p(1, A_{ji}) + p(0, A_{ji})} \right)^{A_{ij}} \left(\frac{p(0, A_{ji})}{p(1, A_{ji}) + p(0, A_{ji})} \right)^{(1-A_{ij})} \quad (\text{A.8})$$

$$P(A_{ji}|A_{ij}) = \left(\frac{p(A_{ij}, 1)}{p(A_{ij}, 1) + p(A_{ij}, 0)} \right)^{A_{ji}} \left(\frac{p(A_{ij}, 0)}{p(A_{ij}, 1) + p(A_{ij}, 0)} \right)^{(1-A_{ji})}. \quad (\text{A.9})$$

In addition to the expected values reported in the article, we can also compute the variances and the covariance between the random variables:

$$\text{Var}[A_{ij}] = \left(\frac{\lambda_{ij}(1 + \eta\lambda_{ji})}{Z_{(ij)}} \right) \left(\frac{1 + \lambda_{ji}}{Z_{(ij)}} \right) \quad (\text{A.10})$$

$$\text{Var}[A_{ji}] = \left(\frac{\lambda_{ji}(1 + \eta\lambda_{ij})}{Z_{(ij)}} \right) \left(\frac{1 + \lambda_{ij}}{Z_{(ij)}} \right) \quad (\text{A.11})$$

$$\text{Cov}[A_{ij}, A_{ji}] = \frac{\eta\lambda_{ij}\lambda_{ji} - \lambda_{ij}\lambda_{ji}}{Z_{(ij)}^2}. \quad (\text{A.12})$$

At each step of the EM algorithm, one updates ρ using equation (3.3) (E-step) and then maximizes $\mathcal{L}(\rho, \Theta)$ with respect to $\Theta = (u, v, w, \eta)$ by setting partial derivatives to zero (M-step). The derivative w.r.t. η is given by:

$$\frac{\partial \mathcal{L}(\rho, \Theta)}{\partial \eta} = \frac{1}{2\eta} \sum_{ij} A_{ij}A_{ji} - \frac{1}{2} \sum_{ij} \frac{\lambda_{ij}\lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta\lambda_{ij}\lambda_{ji} + 1} \stackrel{!}{=} 0, \quad (\text{A.13})$$

that leads to:

$$\eta = \frac{\sum_{ij} A_{ij}A_{ji}}{\sum_{ij} \left[\frac{\lambda_{ij}\lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta\lambda_{ij}\lambda_{ji} + 1} \right]}. \quad (\text{A.14})$$

Similarly, we get the updates for u , v and w :

$$u_{ik} = \frac{\sum_{j,q} A_{ij} \rho_{ijkq}}{\sum_j \left[\frac{\sum_q v_{jq} w_{kq} (1 + \eta \lambda_{ji})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]} \quad (\text{A.15})$$

$$v_{ik} = \frac{\sum_{j,q} A_{ji} \rho_{jikq}}{\sum_j \left[\frac{\sum_q u_{jq} w_{kq} (1 + \eta \lambda_{ij})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]} \quad (\text{A.16})$$

$$w_{kq} = \frac{\sum_{i,j} A_{ij} \rho_{ijkq}}{\sum_{i,j} \left[\frac{u_{ik} v_{jq} (1 + \eta \lambda_{ji})}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}. \quad (\text{A.17})$$

B. Benchmark generative model

The model we propose in the manuscript is able to generate synthetic data with intrinsic community structure and a reciprocity value. It takes as input a set of membership vectors, \mathbf{u}_i and \mathbf{v}_i , affinity matrix w and a pair-interaction parameter η ; the output is a directed and binary network with adjacency matrix A whose pairs of edges are conditionally independent from each other. We use the same formulation as in Safdari *et al.* [10], but our approach differs in that edges between a given pair of nodes are generated stochastically according to the joint probability in equation (2.1), and not according to a two-step sampling procedure. In detail, we assign a value to each pair (A_{ij}, A_{ji}) by considering the vector of cumulative probabilities built using equations (A.1)–(A.4). To enforce sparsity, we multiply λ by a constant ζ , and in order to automatically rescale the expected value in equation (2.7) we have to impose

$$\mathbb{E}[M] = \sum_{ij} \frac{\zeta \lambda_{ij} + \eta \zeta \lambda_{ij} \zeta \lambda_{ji}}{\zeta \lambda_{ij} + \zeta \lambda_{ji} + \eta \zeta \lambda_{ij} \zeta \lambda_{ji} + 1} \quad (\text{B.1})$$

and solve with respect to ζ , where $\mathbb{E}[M]$ is the expected number of edges, a quantity given in input. The benchmark we propose here differs from the one presented in Safdari *et al.* [10] for multiple reasons, as we summarize in Table 1. In addition to those, it is worth mentioning that the competing benchmark in Safdari *et al.* [10] depends on a variable, $cr_{\text{ratio}} = 1 - \eta$, that controls the proportion of edges generated purely by either community or reciprocity effect. This implies that in order to have high reciprocity we may weaken the impact of community effect. This does not happen with the model we propose here, as tie formation can be highly influenced by both reciprocity and community structure at the same time, thus providing a more reliable and truthful representation in certain real world examples.

In the manuscript, we use networks generated with the benchmark proposed above where we fix $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of the pair-interaction parameter η such that we obtain networks with reciprocity values r in the interval $[0, 0.8]$. In detail, we use $\eta \in \{0.1, 10, 20, 40, 80, 140, 280, 500, 1500\}$ to get $r \in \{0, 0.1, 0.2, \dots, 0.8\}$, and we generate 10 different samples for each value of η . Additionally to the data presented in the manuscript, we also report in Appendix C.3 further results on synthetic data generated by varying the average degree $\langle k \rangle$ in the interval $[2, 18]$ while fixing $\eta = 1000$ and the other parameters as above. To generate the membership matrices u and v , we first assign an equal-size unmixed group membership and then we apply the overlapping to 20% of the nodes by drawing those entries from a Dirichlet distribution with parameter $\alpha = 0.1$. The affinity matrix w is generated using an assortative block

structure with main probabilities $p_1 = \langle k \rangle K / N$ and secondary probabilities $p_2 = 0.1 p_1$. Thus, the latent variables $\Theta = (u, v, w, \eta)$ are fixed. Then, edges are drawn according to the generative model described above.

For sake of completeness, we also analysed synthetic networks generated with the model proposed in Safdari *et al.* [10] obtaining similar results and same conclusions. Furthermore, we investigated the behaviour of the models on networks with more than two communities and noticed that results are not highly impacted by this parameter. We do not report them here for sake of brevity.

C. Results on synthetic data

C.1 Edge prediction

We test edge prediction by using a 5-fold cross-validation routine: we divide the dataset into five equal-size groups and train the model on four of them (training set) to infer the parameters; the fifth group is then used as test set to evaluate the existence of edges A_{ij} (in this set). By varying which group we use as test set, we get five trials per realization and the final score is the average over these. To divide the dataset into five folds, we use a symmetric mask, that is, in each trial the training set contains the 80% of the possible entries (A_{ij}, A_{ji}) . In the article, we show the performance of the models in edge prediction when using the marginal and conditional expected values, $\mathbb{E}[A_{ij}]$ and $\mathbb{E}[A_{ij}|A_{ji}]$, respectively. Here, we measure the AUC that is equivalent to the area under the receiver-operating characteristic (ROC) curve [34]. In addition to these results, we can exploit the full joint distribution of our model to answer questions like *what is the probability of jointly observing both edges $i \rightarrow j$ and $j \rightarrow i$?* This is equivalent to assign a value to the pair (A_{ij}, A_{ji}) from the set $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$, that translates the edge prediction task into a classification problem. However, this problem becomes trivial if the model predicts all entries equal to $(0, 0)$: in this case, we will get high performance just because of the high sparsity of the data. For this reason, we compute the accuracy only for entries in the test set that have at least one edge. For those, we predict the label associated to the highest probability among $[p_{01}, p_{10}, p_{11}]$, where these are computed by using equations (A.1)–(A.3) with the inferred parameters. We then compute the accuracy between true and predicted labels, where a value equal to 1 means perfect recovery. As baselines, we use a uniform random probability (RP) over the number of possible labels in the training set, and the accuracy obtained by using as prediction the label with the highest relative frequency in the training set (MRF). The results are shown in Fig. C1, where we can observe a V-shape. Reciprocity equal to zero ($r = 0$) means the networks have no reciprocated edges, and higher its value higher the frequency of the label $(1, 1)$. Thus, in the regime $0 \leq r \leq 0.5$ the performance decreases because the problem becomes more difficult by reaching the point where labels have similar relative frequencies (MRF \approx RP when $r = 0.5$). In this scenario, JointCRep outperforms the baselines with a bigger gap as the reciprocity increases. When $r > 0.5$ the problem becomes easier due to the increasing proportion of the label $(1, 1)$. Here, predicting all entries equal to $(1, 1)$ results in higher performance (MRF). However, this is another trivial situation that should be ignored when analysing the performance in edge prediction tasks.

C.2 Reproducing network topological properties

Figure C2 shows the performance of each model in reproducing the average degree in sampled networks. While JointCRep and MT recover this feature despite the different values of reciprocity, CRep produces samples with a lower average degree than the one given in input as r increases. This happens because, in high reciprocity settings, CRep produces sampled networks with fewer edges but higher weights. Hence,

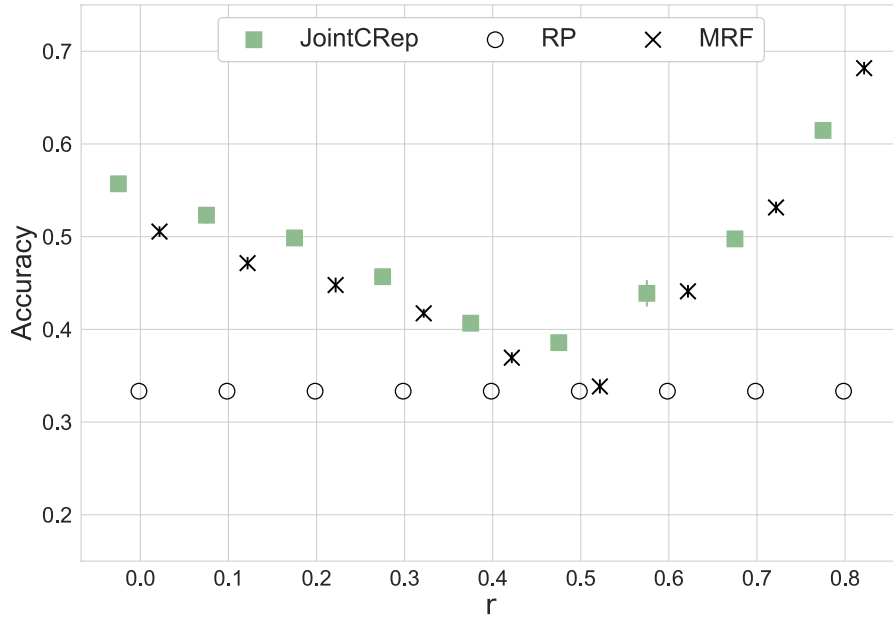


FIG. C1. Edge prediction with joint distributions in synthetic networks. Synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of reciprocity r . Results are averages and standard deviations over 10 synthetic networks and over 5-folds of cross-validation test sets. Edge prediction performance is measured with accuracy, and as baselines, we consider the uniform random probability (RP) and the maximum relative frequency (MRF).

CONTISCIANI *ET AL.*

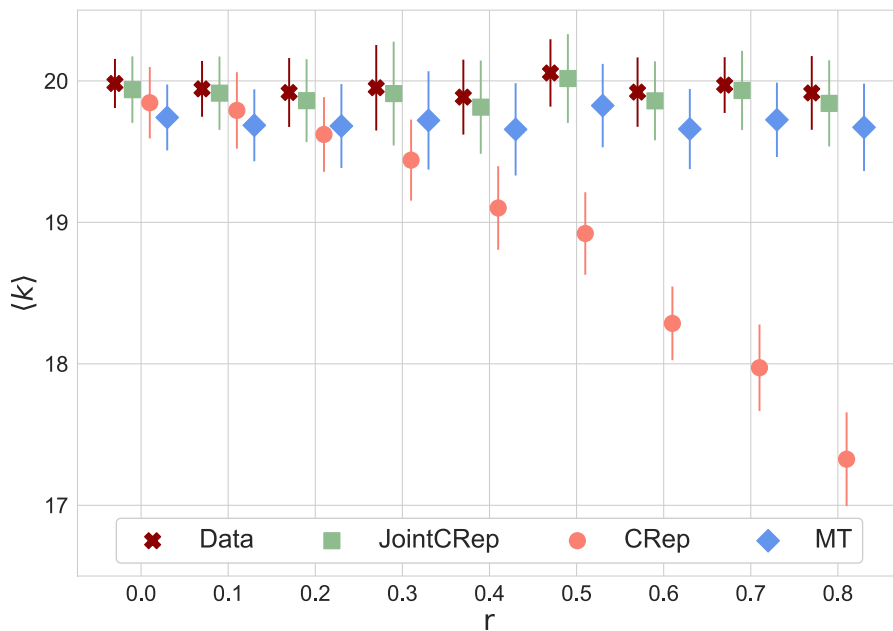


FIG. C2. Average degree in sampled synthetic networks. Synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, $\langle k \rangle = 20$ average degree and different values of reciprocity r . Results are empirical averages and standard deviations over 50 samples of 10 independent synthetic networks (five samples per input network). We measure the average degree and the cross (x) markers indicate the average on 10 input networks.

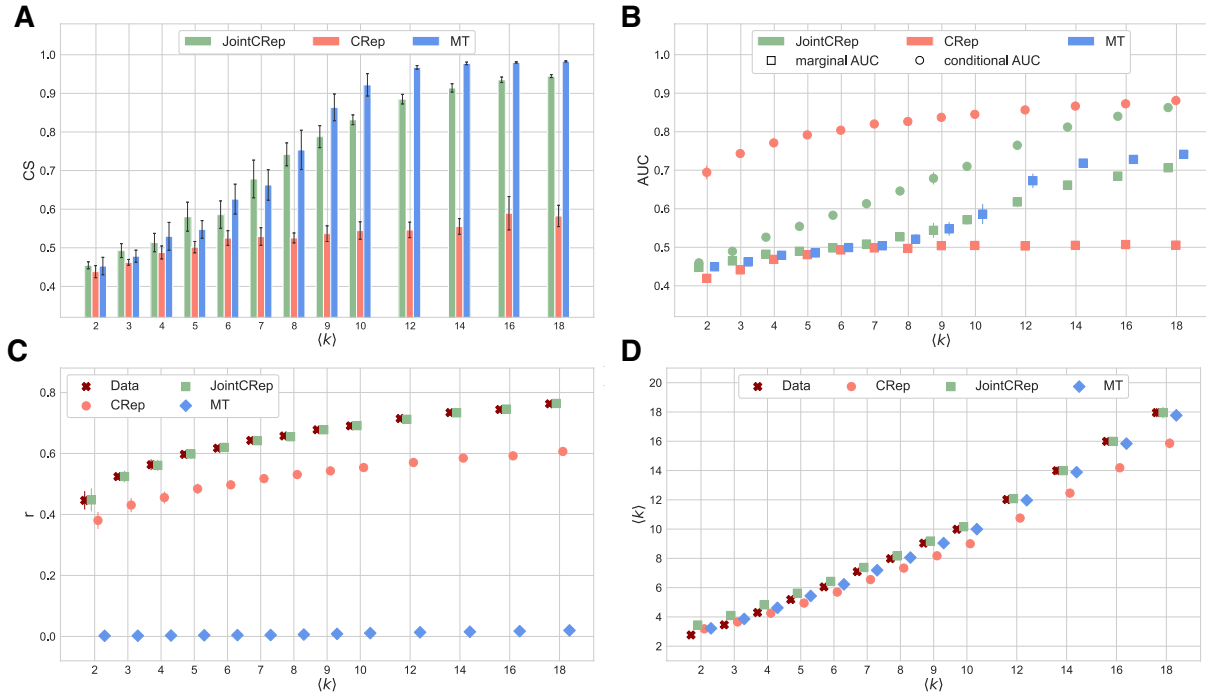


FIG. C3. Results on synthetic networks with different average degrees. Synthetic networks with $N = 1000$ nodes, $K = 2$ overlapping communities, pair-interaction parameter $\eta = 1000$, and different values of average degree $\langle k \rangle$. (A,B) Results are averages and standard deviations over 10 synthetic networks of (A) cosine similarity and (B) AUC. The latter measures the edge prediction performance over 5-folds of cross-validation test sets, and the baseline is the random value 0.5. (C,D) Results are empirical averages and standard deviations over 50 samples of 10 independent synthetic networks (five samples per input network). We measure (C) the reciprocity and (D) the average degree, and the cross (x) markers indicate the average on 10 input networks.

while the average degree decreases, the weighted average degree better reflects the input feature (not shown here).

C.3 Analysis on synthetic data with different average degrees

In addition to the results provided in the manuscript, we also analyse synthetic networks with different values of average degree. Figure C3 shows the performance of the models in community detection and edge prediction tasks, as well as in reproducing topological properties in sampled networks. Similar to the results in the manuscript, JointCRep follows the behaviour of MT both in terms of CS and marginal AUC, for which performance improves as the average degree increases, as expected for community detection-only methods. On the other hand, CRep is only partially affected by the different values of average degree, as also shown in [10]. The plots highlight that the strength of CRep is not retrieving communities, rather its ability to predict missing edges by using the conditional distribution, regardless the average degree. In Fig. C3, we can also notice that even though JointCRep is affected by the average degree, its conditional AUC improves the marginal AUCs already when $\langle k \rangle = 4$, and it reaches good values from $\langle k \rangle = 10$. Moreover, JointCRep outperforms the other methods in recovering reciprocity in sampled networks across different values of average degree. Overall, these results suggest that JointCRep is a valuable tool also in networks with low-medium average degree providing good communities, reasonable edge predictions, and sampled networks with topological features that resemble that of the real data.

D. Results on real-world datasets

D.1 Loss functions

In addition to the AUC, we use the Log Loss (or Binary Cross-Entropy) and the L1 Loss (or Mean Absolute Error) to measure the performance of the methods in edge prediction and network reconstruction. The Log Loss for binary classification is defined as

$$-\frac{1}{M} \sum_{ij} [A_{ij} \log P(A_{ij}) + (1 - A_{ij}) \log (1 - P(A_{ij}))], \quad (\text{D.1})$$

where A_{ij} indicates the entry of the adjacency matrix, $P(A_{ij})$ denotes the probability of the existence of the edge, and M is the total number of edges.

Instead, the L1 Loss is given by

$$\frac{1}{M} \sum_{ij} |A_{ij} - P(A_{ij})|. \quad (\text{D.2})$$

For both metrics, lower values indicate better performance and a loss of 0 denotes perfect predictions. While the Log Loss does not have an upper bound, the L1 Loss is equal to 1 in the worst-case scenario of predicting every existing edge with probability $P(A_{ij} = 1) = 0$ and every non-existing edge with probability $P(A_{ij} = 1) = 1$. Moreover, they differ in the extent to which they penalize mistakes: the Log Loss is more sensitive to large disagreements between true and predicted values than the L1 Loss. This means that the Log Loss prefers predictions with more mistakes of low magnitude than predictions with fewer mistakes but of larger magnitude.

D.2 Analysis of a high-school social network

Table D1 displays the results for the edge prediction task in the high-school social network. CRep performs the best both in terms of AUC and Log Loss when using the conditional probabilities. On the other hand, JointCRep is the best when considering the L1 Loss. This is explained by the behaviour of our model that tends to predict fewer edges with more intensity, differently to the other models which predict more edges with low-medium probabilities. As a remark, the dataset presents an average degree $\langle k \rangle = 6.45$ and it is highly sparse. This feature makes this task hard because there is only little information in input when considering 5-fold cross-validation splits, and some folds may result in unreliable results. Nevertheless, results show that all models are performing reasonably well at this task given this sparse regime.

Comparing the communities inferred by the various methods, Table D2 shows the overlapping modularity obtained for the partitions of Fig. 4 for various aggregation functions. Table D3 reports the penalties for the network reconstruction task, showing that JointCRep has best performance in terms of both Log Loss and L1 Loss. Finally, Table D4 shows the topological properties in the high-school social network and its sampled networks, showing how JointCRep achieves on average values that are more similar to those observed on the input data.

D.3 Analysis of a vampire bat network

Table D5 shows results for edge prediction tasks using a 5-fold cross-validation routine. Similar to the high-school dataset, all the models obtain good performance given the sparse regime, although values

TABLE D1 *Edge prediction in the high-school social network. Results are averages and standard deviations over 5-folds of cross-validation test sets. Edge performance is measured with three different metrics \mathcal{F} : AUC, Log Loss and L1 Loss. The AUC measures the probability that a randomly selected edge has higher expected value than a randomly selected non-existing edge, and the baseline is the random value 0.5. The Log Loss and the L1 Loss are penalty measures defined in Appendix D.1, that quantify the difference between two input networks by taking into account the probability of the existence of an edge and computing a penalty for each mistake in predicting the observed value. The metrics are computed by using either the marginal probability $P(A_{ij}|\Theta)$ or the conditional probability $P(A_{ij}|A_{ji}, \Theta)$. Note that the last is not available for MT because the conditional and marginal distributions coincide. The best performance for each metric is in bold*

\mathcal{F}	$P(A_{ij} \Theta)$			$P(A_{ij} A_{ji}, \Theta)$	
	JointCRep	CRep	MT	JointCRep	CRep
AUC	0.610 \pm 0.061	0.650 \pm 0.109	0.668 \pm 0.111	0.626 \pm 0.073	0.786 \pm 0.055
Log Loss	0.825 \pm 0.191	0.678 \pm 0.190	0.726 \pm 0.284	0.820 \pm 0.213	0.492 \pm 0.124
L1 Loss	0.133 \pm 0.014	0.139 \pm 0.015	0.132 \pm 0.026	0.122 \pm 0.019	0.125 \pm 0.014

TABLE D2 *Modularity for the high-school social network. Values are computed using the overlapping formulation as in Nicosia et al. [28], and \mathcal{F} denotes the aggregation function considered in each row. We use the mixed-membership partitions determined by the matrix u inferred by JointCRep, CRep and MT. Results are similar for the matrix v*

\mathcal{F}	JointCRep	CRep	MT
Mean	0.74	0.72	0.75
Max	0.65	0.62	0.48
Product	0.55	0.53	0.73

TABLE D3 *High-school network reconstruction: comparison between true and reconstructed networks. \mathcal{F} denotes the function considered in each row, defined in Appendix D.1. The metrics are computed by using either the marginal probability $P(A_{ij}|\Theta)$ or the conditional probability $P(A_{ij}|A_{ji}, \Theta)$ of each method with the inferred parameters. Note that the last is not available for MT because the conditional and marginal distributions coincide. The best performance for each metric is in bold*

\mathcal{F}	$P(A_{ij} \Theta)$			$P(A_{ij} A_{ji}, \Theta)$	
	JointCRep	CRep	MT	JointCRep	CRep
Log Loss	0.144	0.307	0.165	0.128	0.185
L1 Loss	0.093	0.137	0.106	0.077	0.120

TABLE D4 *Topological properties in the high-school social network and its sampled networks. Results are averages and standard deviations over five samples. We measure the number of nodes N , the number of edges M , the average degree $\langle k \rangle$, the reciprocity r and the clustering coefficient cc*

	N	M	$\langle k \rangle$	r	cc
Data	31	100	6.45	0.52	0.38
JointCRep	30.8 ± 0.45	90.8 ± 6.76	5.89 ± 0.38	0.47 ± 0.06	0.20 ± 0.02
CRep	30.6 ± 0.55	77.8 ± 12.79	5.08 ± 0.81	0.49 ± 0.08	0.11 ± 0.05
MT	31 ± 0	79.8 ± 2.05	5.15 ± 0.13	0.21 ± 0.04	0.24 ± 0.04

TABLE D5 *Edge prediction in the vampire bat network. Results are averages and standard deviations over 5-folds of cross-validation test sets. Edge performance is measured with three different metrics \mathcal{F} : AUC, Log Loss and L1 Loss. The AUC measures the probability that a randomly selected edge has higher expected value than a randomly selected non-existing edge, and the baseline is the random value 0.5. The Log Loss and the L1 Loss are penalty measures defined in Appendix D.1, that quantify the difference between two input networks by taking into account the probability of the existence of an edge and computing a penalty for each mistake in predicting the observed value. The metrics are computed by using either the marginal probability $P(A_{ij}|\Theta)$ or the conditional probability $P(A_{ij}|A_{ji}, \Theta)$. Note that the last is not available for MT because the conditional and marginal distributions coincide. The best performance for each metric is in bold*

\mathcal{F}	$P(A_{ij} \Theta)$			$P(A_{ij} A_{ji}, \Theta)$	
	JointCRep	CRep	MT	JointCRep	CRep
AUC	0.687 ± 0.078	0.627 ± 0.079	0.629 ± 0.073	0.715 ± 0.098	0.772 ± 0.063
Log Loss	1.514 ± 0.282	0.961 ± 0.196	1.804 ± 0.147	1.391 ± 0.269	0.812 ± 0.261
L1 Loss	0.277 ± 0.031	0.340 ± 0.014	0.291 ± 0.020	0.229 ± 0.021	0.296 ± 0.008

TABLE D6 *Vampire bat network reconstruction: comparison between true and reconstructed networks. \mathcal{F} denotes the function considered in each row, defined in Appendix D.1. The metrics are computed by using either the marginal probability $P(A_{ij}|\Theta)$ or the conditional probability $P(A_{ij}|A_{ji}, \Theta)$ of each method with the inferred parameters. Note that the last is not available for MT because the conditional and marginal distributions coincide. The best performance for each metric is in bold*

\mathcal{F}	$P(A_{ij} \Theta)$			$P(A_{ij} A_{ji}, \Theta)$	
	JointCRep	CRep	MT	JointCRep	CRep
Log Loss	0.173	0.466	0.302	0.159	0.414
L1 Loss	0.110	0.308	0.207	0.103	0.275

were slightly better in the high-school case. Also in this case, JointCRep achieves best results in terms of L1 Loss, and CRep is more robust in terms of Log Loss. Table D6 reports the quantitative results for the network reconstruction of the vampire bat dataset, in terms of Log and L1 Loss. Here, JointCRep outperforms the other methods as also shown in Fig. 7.

REFERENCES

1. FELL, D. A. & WAGNER, A. (2000) The small world of metabolism. *Nat. Biotechnol.*, **18**, 1121–1122.
2. NEWMAN, M. E. (2001) The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, **98**, 404–409.
3. WATTS, D. J. & STROGATZ, S. H. (1998) Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440–442.
4. WILLIAMS, R. J. & MARTINEZ, N. D. (2000) Simple rules yield complex food webs. *Nature*, **404**, 180–183.
5. GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. & AIROLDI, E. M. (2010) A survey of statistical network models. *Found. Trends Mach. Learn.*, **2**, 129–233.
6. FORTUNATO, S. (2010) Community detection in graphs. *Phys Rep.*, **486**, 75–174.
7. BALL, B., KARRER, B. & NEWMAN, M. E. (2011) Efficient and principled method for detecting communities in networks. *Phys. Rev. E*, **84**, 036103.
8. DE BACCO, C., POWER, E. A., LARREMORE, D. B. & MOORE, C. (2017) Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E*, **95**, 042317.
9. PEIXOTO, T. P. (2022) Disentangling homophily, community structure, and triadic closure in networks. *Phys. Rev. X*, **12**, 011004.
10. SAFDARI, H., CONTISCIANI, M. & DE BACCO, C. (2021) Generative model for reciprocity and community detection in networks. *Phys. Rev. Res.*, **3**, 023209.
11. SESHADHRI, C., SHARMA, A., STOLMAN, A. & GOEL, A. (2020) The impossibility of low-rank representations for triangle-rich complex networks. *Proc. Natl. Acad. Sci. USA*, **117**, 5631–5637.
12. HOLLAND, P. W., LASKEY, K. B. & LEINHARDT, S. (1983) Stochastic blockmodels: first steps. *Soc. Netw.*, **5**, 109–137.
13. KARRER, B. & NEWMAN, M. E. (2011) Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, **83**, 016107.
14. WASSERMAN, S., FAUST, K. ET AL. (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
15. DE BACCO, C., CONTISCIANI, M., CARDOSO-SILVA, J., SAFDARI, H., BAPTISTA, D., SWEET, T., YOUNG, J.-G., KOSTER, J., ROSS, C. T., McELREATH, R. ET AL. (2021) Latent network models to account for noisy, multiply-reported social network data. *arXiv preprint arXiv:2112.11396*.
16. READY, E. & POWER, E. A. (2021) Measuring reciprocity: double sampling, concordance, and network construction. *Netw. Sci.*, **9**, 387–402.
17. LI, W., ASTE, T., CACCIOLI, F. & LIVAN, G. (2019) Reciprocity and impact in academic careers. *EPJ Data Sci.*, **8**, 20.
18. GARLASCHELLI, D. & LOFFREDO, M. I. (2004) Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.*, **93**, 268701.
19. NEWMAN, M. E., FORREST, S. & BALTHROP, J. (2002) Email networks and the spread of computer viruses. *Phys. Rev. E*, **66**, 035101.
20. HOLLAND, P. W. & LEINHARDT, S. (1981) An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.*, **76**, 33–50.
21. PARK, J. & NEWMAN, M. E. (2004) Statistical mechanics of networks. *Phys. Rev. E*, **70**, 066117.
22. ROBINS, G., PATTISON, P., KALISH, Y. & LUSHER, D. (2007) An introduction to exponential random graph (p*) models for social networks. *Soc. Netw.*, **29**, 173–191.
23. SNIJDERS, T. A., PATTISON, P. E., ROBINS, G. L. & HANDCOCK, M. S. (2006) New specifications for exponential random graph models. *Sociol. Methodol.*, **36**, 99–153.
24. WASSERMAN, S. & ANDERSON, C. (1987) Stochastic a posteriori blockmodels: construction and assessment. *Soc. Netw.*, **9**, 1–36.
25. ISING, E. (1925) Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, **31**, 253–258.
26. DAI, B., DING, S., WAHBA, G. ET AL. (2013) Multivariate Bernoulli distribution. *Bernoulli*, **19**, 1465–1483.
27. COLEMAN, J. S. (1964) *Introduction to Mathematical Sociology*. Glencoe: London Free Press.

28. NICOSIA, V., MANGIONI, G., CARCHIOLO, V. & MALGERI, M. (2009) Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech.*, **2009**, P03024.
29. FORTUNATO, S. & BARTHELEMY, M. (2007) Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA*, **104**, 36–41.
30. NEWMAN, M. E. (2016) Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E*, **94**, 052315.
31. CARTER, G. G. & WILKINSON, G. S. (2013) Food sharing in vampire bats: reciprocal help predicts donations more than relatedness or harassment. *Proc. R. Soc. B*, **280**, 20122573.
32. SAFDARI, H., CONTISCIANI, M. & DE BACCO, C. (2022) Reciprocity, community detection, and link prediction in dynamic networks. *J. Phys.*, **3**, 015010.
33. CONTISCIANI, M., POWER, E. A. & DE BACCO, C. (2020) Community detection with node attributes in multilayer networks. *Sci. Rep.*, **10**, 1–16.
34. HANLEY, J. A. & MCNEIL, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

OPEN ACCESS



PAPER

Reciprocity, community detection, and link prediction in dynamic networks

RECEIVED

26 August 2021

REVISED

18 January 2022

ACCEPTED FOR PUBLICATION

8 February 2022

PUBLISHED

28 February 2022

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Hadiseh Safdari , Martina Contisciani and Caterina De Bacco*

Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen 72076, Germany

* Author to whom any correspondence should be addressed.

E-mail: hadiseh.safdari@tuebingen.mpg.de, martina.contisciani@tuebingen.mpg.de and caterina.debacco@tuebingen.mpg.de**Keywords:** community detection, networks, probabilistic inference, dynamical networks**Abstract**

Many complex systems change their structure over time, in these cases dynamic networks can provide a richer representation of such phenomena. As a consequence, many inference methods have been generalized to the dynamic case with the aim to model dynamic interactions. Particular interest has been devoted to extend the stochastic block model and its variant, to capture community structure as the network changes in time. While these models assume that edge formation depends only on the community memberships, recent work for static networks show the importance to include additional parameters capturing structural properties, as reciprocity for instance. Remarkably, these models are capable of generating more realistic network representations than those that only consider community membership. To this aim, we present a probabilistic generative model with hidden variables that integrates reciprocity and communities as structural information of networks that evolve in time. The model assumes a fundamental order in observing reciprocal data, that is an edge is observed, conditional on its reciprocated edge in the past. We deploy a Markovian approach to construct the network's transition matrix between time steps and parameters' inference is performed with an expectation-maximization algorithm that leads to high computational efficiency because it exploits the sparsity of the dataset. We test the performance of the model on synthetic dynamical networks, as well as on real networks of citations and email datasets. We show that our model captures the reciprocity of real networks better than standard models with only community structure, while performing well at link prediction tasks.

1. Introduction

Many real networks are dynamical, i.e., the pattern of interactions between their nodes vary over time, e.g., network of exchanged emails in a company. The abundance of such datasets and the development of optimal numerical methods have led to a growing number of studies in this field [1–4]. In addition, interactions between nodes can be reciprocated, e.g., the people whom one retweets and the number of times she retweets them vary over time; so do the papers that researchers cite in their manuscripts and papers that cite one's scientific output. This latter issue has received little attention in previous studies.

Among the main approaches to study these systems, latent variable models assume that the existence of an edge between any pair of nodes is independent of other nodes, and is conditional on some latent variables which incorporate the hidden structure of the network. These techniques mainly focus on community membership as the main relevant latent variable, e.g., in the case of citations, the people who cite each other's works, inadvertently form a community. The stochastic block model (SBM) [5–7] and its variants provide flexible network generative models [8, 9]. In this framework, nodes are initially partitioned into communities, then edges are created between nodes, based on their community membership. There are several variants of dynamical equivalents of stochastic block model (DSBM) [10–14] which capture transition of community membership over time, reflecting the evolution of edge formation. Peixoto and Rosvall [15], and Matias *et al* [16] develop a

non-parametric temporal SBM. Gauvin *et al* [17] consider non-negative tensor factorization, where communities are static but the affinity matrix changes over time. Bovev *et al* [18] use flow of random walkers co-evolving in the dynamic network to define communities. Various methods have been used to address whether the community membership or connectivity parameters could change over time, see [19] for a review. For instance, one could assume that communities are fixed in time but the connectivity parameters across groups changes, as in [11, 17], or that communities change in time [10, 20–22].

In Zhang *et al* [12], the authors extend some of the popular methods of modeling network structure, e.g., SBM, to represent dynamic networks. The main idea behind their Markovian approach is to find transition rates of appearance and disappearance of edges over time. Based on these rates, they were able to calculate the average probability of edges over all time steps, hence, they estimate a steady state probability distribution for each network model, depending on its structural parameters. Although the approach followed in [12] is efficient and analytically grounded, it was developed for models that incorporate communities as the only latent variable.

Nevertheless, in directed real-world networks, community membership may not be the only factor influencing network structure. Reciprocity, i.e., the tendency of a pair of nodes to form edges on both directions, has been subject of many studies [23–25] as a crucial factor to determine the structure of networks, in particular in social networks. Bartolucci *et al* [26] assume local conditional independence between pairs of edges, i.e., dyads, and extend the SBM to account for the reciprocal patterns in directed dynamical networks. Furthermore, they established various specifications of the proposed model corresponding to different reciprocal assumptions.

Recently, a generative model (CRep) has been introduced that, in addition to community membership, includes reciprocity as latent variable that dictates formation of edges between the nodes [25]. In other words, the appearance of a directed edge from node i to j not only depends on the community that the nodes belong to, but also is affected by the existence of the edge from j to i . In the case of citation network, it is more likely for an author to cite those other who already cited her, implying overlapping research areas.

In this work, following the approach in [12], we extend CRep and propose a continuous-time Markov process model for dynamic networks (DynCRep). Observing the system at discrete points in time, at each time step the transition rates of appearance and disappearance of a directed edge between two nodes depends on the current community membership of the nodes, as well as on the existence of a reciprocated edge between them.

We validate the applicability of the proposed model and its inference approach by performing experiments on real and synthetic networks for community detection and link prediction. We apply the model to synthetic datasets and observe that DynCRep shows a reasonable performance in terms of link prediction. Moreover, we test the model performance on real-world datasets in the domain of social and online communication to reproduce reciprocity, with promising results.

2. Model

In our model, the temporally evolving network is captured in snapshots taken at fixed intervals, from $t = 0$ to $T + 1$. $A(t)$ represents the dynamic adjacency matrix of the network, where a non-zero value of $A_{ij}(t)$ represents a weighted edge from i to node j at time t , and $A_{ij}(t) = 0$ denotes no interaction. We assume that the total number of nodes is fixed over time, i.e., new nodes do not enter into the network, and nodes do not leave it; instead, existing edges can appear and disappear. We focus on directed, and weighted networks.

A matrix $w(t)$ of dimensions $K \times K$ determines the evolving structure of the K communities over time and we refer to $w(t)$ as the affinity matrix. Different assumptions about $w(t)$ result in communities with different structures. For instance, in the case of diagonal entries being greater than off-diagonal ones, communities are assortative—that is, individuals are more inclined towards intra-community interactions than inter-community interactions. The K -dimensional vectors $u_i(t)$ and $v_i(t)$ denote the out-going and in-coming communities at time step t , respectively.

Here, we keep the community membership constant over time; hence, we drop the notion of time dependency. We develop the model in two different varieties: (1) the affinity matrix varies over time (w-DYN), i.e., the connectivity pattern between communities changes over time, for instance, a group of nodes which form a community at time step t could be peripheral nodes at another time step [11], and (2) the affinity matrix also remains static (w-STATIC).

Following the continuous-time Markov process approach in [12], we assume that networks evolve on the real-valued times; hence, the appearance and disappearance of the edges are continuous parameters. However, we observe the network at discrete time steps. At each time step, a Poisson distribution governs the existence of edges between nodes such that an edge between two nodes is formed at a rate $\hat{\lambda}_{ij}(t)$. This rate depends on both the community that nodes belong to, and the existence of the reciprocated tie at the previous

time step:

$$\begin{aligned}\hat{\lambda}_{ij}(t) &= \lambda_{ij}(t) + \eta A_{ji}(t-1) \\ &\equiv \sum_{k,q} u_{ik} v_{jq} w_{kq}(t) + \eta A_{ji}(t-1),\end{aligned}\quad (1)$$

where η as a hyperparameter regulates the reciprocity effects, similarly as in [25]. The difference between equation (1) and the edge probability in [25] is that the dependency on the reciprocated tie is on the previous time step, while standard CRep considers only the same time t , being an approach valid for static networks. Furthermore, an edge could disappear with rate μ .

2.1. Dynamic CRep

The aim of this study is to infer the latent parameters of the model, namely, $\Theta \equiv \{u, v, w, \eta, \mu\}$, given the adjacency matrix observed at each time step. To this end, we perform this inference task by maximizing the log-likelihood. Given Θ , all the pairs of nodes are conditionally independent; as a result, the joint-probability of the node-pairs could be approximated by a factorized form. Here, we develop a Markov process, according to which, at every time step, the probability of edges depends only on the previous time step:

$$\begin{aligned}P(\{A(t)\}|\Theta) &= P(\{A(t)\}|\{A(t-1)\}, \Theta) \\ &= \prod_{ij} \left\{ P(A_{ij}(t)|A_{ji}(0), \Theta) \prod_{t=1}^T \{P(A_{ij}(t)|A_{ij}(t-1), A_{ji}(t-1), \Theta)\} \right\}.\end{aligned}\quad (2)$$

We further assume that at the initial time step the probability $A_{ij}(0)$ of an edge between two nodes follows a Poisson distribution with mean $\hat{\lambda}_{ij} = \lambda_{ij}(0)$, i.e., there is no reciprocated edge in the past:

$$P(A_{ij}(0)|A_{ji}(0), \Theta) = \frac{e^{-\lambda_{ij}(0)} \lambda_{ij}(0)^{A_{ij}(0)}}{A_{ij}(0)!}.\quad (3)$$

At each time-step, edges appear with rate $\hat{\lambda}_{ij}(t)$, and disappear with rate μ . We follow an approach similar to that of Zhang *et al* [12] and calculate the probability of the existence of edges by solving a master equation. Defining $p_{ij}^k(t)$ as the probability of having k edges, i.e., an edge with the weight equal to k , between nodes i, j at time t , this quantity satisfies the following master equation:

$$\frac{dp_{ij}^k(t)}{dt} = \hat{\lambda}_{ij}(t) p_{ij}^{k-1}(t) + (k+1)\mu p_{ij}^{k+1}(t) - (\hat{\lambda}_{ij}(t) + k\mu) p_{ij}^k(t).\quad (4)$$

To solve this equation, we use a generating function approach [27], by defining $g(z, t) = \sum_{k=0}^{\infty} p^k(t) z^k$. The solution for the generating function,

$$g(z, t) = f \left[(z-1)e^{-\mu t} \right] e^{\frac{(z-1)\hat{\lambda}_{ij}(t)}{\mu}},\quad (5)$$

could be expanded in terms of z to give us p_{ij}^t (more details in section S1 (<https://stacks.iop.org/JPCOMPLEX/03/015010/mmedia>)). There are four possible transitions from time $t-1$ to t : (1) there is no edge neither at time $t-1$, nor at t ; (2) the appearance of an edge from non-edge, (3) disappearance of an existing edge, and (4) an existing edge remains; with the following probabilities, respectively,

$$\begin{aligned}p_{0 \rightarrow 0} &= e^{-\beta(\lambda_{ij}(t) + \eta A_{ji}(t))} \\ p_{0 \rightarrow 1} &= \beta(\lambda_{ij}(t) + \eta A_{ji}(t)) e^{-\beta(\lambda_{ij}(t) + \eta A_{ji}(t))} \\ p_{1 \rightarrow 0} &= \beta e^{-\beta(\lambda_{ij}(t) + \eta A_{ji}(t))} \\ p_{1 \rightarrow 1} &= (1 - \beta) e^{-\beta(\lambda_{ij}(t) + \eta A_{ji}(t))},\end{aligned}\quad (6)$$

where $\beta = 1 - e^{-\mu}$. This leads to the following time-dependent, log-likelihood:

$$\begin{aligned}
 L(T, \Theta) &= \log[P(\{A(t)\}|\{A(t-1)\}, \Theta)] \\
 &= \sum_{ij} \left\{ \log \left[e^{-\lambda_{ij}(t)} \lambda_{ij}(t)^{A_{ij}(0)} \right] + \sum_{t=1}^T \log \left[e^{-\beta(\lambda_{ij}(t) + \eta A_{ji}(t))} \right. \right. \\
 &\quad \left. \left. \times [\beta (\lambda_{ij}(t) + \eta A_{ji}(t))]^{(1-A_{ij}(t-1))A_{ij}(t)} \beta^{A_{ij}(t-1)(1-A_{ij}(t))} \times (1-\beta)^{A_{ij}(t-1)A_{ij}(t)} \right] \right\}. \quad (7)
 \end{aligned}$$

We add parameters' regularization by assuming gamma-distributed priors for the membership vectors:

$$P(u_{ik}; a, b) \propto u_{ik}^{a-1} e^{-bu_{ik}}, \quad (8)$$

where $a \geq 1$, to ensure the maximization of the log-likelihood (the second derivative must be negative), similarly for the v_{ik} . This adds new terms to the log-likelihood:

$$\mathcal{L}(T, \Theta) = L(T, \Theta) + (a-1) \sum_{i,k} \log u_{ik} - b \sum_{i,k} u_{ik} + (a-1) \sum_{i,k} \log v_{ik} - b \sum_{i,k} v_{ik}. \quad (9)$$

In the experiments below we set the values of the hyper-priors to enforce sparsity, i.e., $a = 1.5, b = 10$.

Maximizing $\mathcal{L}(T, \Theta)$ requires taking the derivative of equation (9) w.r.t. each parameter individually and setting them to zero. Because the summations in the logarithm render the calculations difficult, we employ a variational approximation using Jensen's inequality. Inference is then performed using the expectation-maximization algorithm (EM); details are provided in section S1A.

Hitherto, we have included all the dependencies on the reciprocated edge $A_{ji}(t-1)$ by considering the previous time step $t-1$. However, the model still applies if we incorporate the reciprocated edge at the same time step, i.e., considering $A_{ji}(t)$. This choice may depend on the application itself based on the expectations and insight of the practitioner from the reciprocity effects. Alternatively, one can choose between these two options with model-selection criteria. In our experiments on real data we deployed them both, and presented the version that performs best in cross-validation tasks (section S5A).

We continue with two specifications of the model with different assumptions on the temporal evolution of the affinity matrix. In the first approach, w-DYN, the affinity matrix is treated as a time-dependent variable; while the community membership vectors, u_i, v_i , are kept static over time. Notice that a similar scenario could be obtained by fixing w and changing u_i, v_i in time [11], our model can be easily adapted to accommodate this alternative interpretation. Our model assumes fixed number of communities K . As we consider a mixed-membership model, we have the flexibility of allowing nodes to belong to various communities and with various intensities, thus allowing to capture the likelihood of the data well by effectively changing how an entry u_{ik} or v_{ik} impacts the magnitude of $\lambda_{ij}(t)$ via $w(t)$ in the w-DYN scenario, while keeping K constant.

In the second scenario, w-STATIC, the affinity matrix is kept static as well. The purpose of considering these scenarios is to make the model flexible in dealing with various community structures (see sections S2 to S4 for more details on each scenarios). Notice that in the case of w-STATIC, although all the latent variables are fixed in time, the network can still evolve, as edges appear and disappear based on the parameters β and μ . This is also the case for the Markov model (without reciprocity) in [12].

For instance, the EM algorithm for w-STATIC yields:

$$u_{ik} = \frac{a - 1 + \sum_{j,q,t} \rho_{ij}^{(1)}(t) \phi_{ijkq} \hat{A}_{ij}(t)}{b + \sum_{j,q} v_{jq} w_{kq} (1 + \beta T)} \quad (10)$$

$$v_{jq} = \frac{a - 1 + \sum_{i,k,t} \rho_{ij}^{(1)}(t) \phi_{ijkq} \hat{A}_{ij}(t)}{b + \sum_{i,k} u_{ik} w_{kq} (1 + \beta T)} \quad (11)$$

$$w_{kq} = \frac{\sum_{i,j,t} \rho_{ij}^{(1)}(t) \phi_{ijkq} \hat{A}_{ij}(t)}{\sum_{ij} u_{ik} v_{jq} (1 + \beta T)} \quad (12)$$

$$\eta = \frac{\sum_{i,j,t} \rho_{ij}^{(2)}(t) \hat{A}_{ij}(t)}{\sum_{ij} \sum_{t=1}^T \beta A_{ji}(t-1)}, \quad (13)$$

Algorithm 1. DynCRep (w-DYN): EM algorithm.

Input: network $A(t) = \{A_{ij}(t)\}_{i,j=1}^N$,
 number of communities K .

Output: membership $u = [u_{ik}]$, $v = [v_{ik}]$; network
 affinity matrix $w(t) = [w_{kq}(t)]$; reciprocity
 parameter η ; edge disappearance rate $\beta(t)$.

Initialize $u, v, w(t), \eta, \beta(t)$ at random.
 Repeat until \mathcal{L} converges:

- Calculate $\rho_1(t)$ and $\phi(t)$ (E-step):

$$\rho_{ij}^{(1)}(t) = \frac{\lambda_{ij}(t)}{\lambda_{ij}(t) + \eta A_{ji}(t)}, \quad \rho_{ij}^{(2)}(t) = \frac{\eta A_{ji}(t)}{\lambda_{ij}(t) + \eta A_{ji}(t)},$$

$$\phi_{ijkq}(t) = \frac{u_{ik} v_{jq} w_{kq}(t)}{\sum_{k,q} u_{ik} v_{jq} w_{kq}(t)}.$$
- Update parameters Θ (M-step):
 - For each node i and community k update memberships:

$$u_{ik} = \frac{a - 1 + \sum_{j,q,t} \rho_{ij}^{(1)}(t) \phi_{ijkq}(t) \hat{A}_{ij}(t)}{b + \sum_{j,q} v_{jq} \sum_{t=0}^T \hat{\beta}(t) w_{kq}(t)}$$

$$v_{ik} = \frac{a - 1 + \sum_{j,q,t} \rho_{ij}^{(2)}(t) \phi_{ijkq}(t) \hat{A}_{ij}(t)}{b + \sum_{j,q} u_{jq} \sum_{t=0}^T \hat{\beta}(t) w_{kq}(t)}$$
 - For each pair (k, q) update affinity matrix:

$$w_{kq}(t) = \frac{\sum_{i,j} \rho_{ij}^{(1)}(t) \phi_{ijkq}(t) \hat{A}_{ij}(t)}{\sum_{i,j} u_{ik} v_{jq} \hat{\beta}(t)}$$
 - Update reciprocity parameter:

$$\eta = \frac{\sum_{i,j,t} \rho_{ij}^{(2)}(t) \hat{A}_{ij}(t)}{\sum_{i,j,t=1} \hat{\beta}(t) A_{ji}(t-1)}$$

where we defined $\hat{A}_{ij}(t) = A_{ij}(t)(1 - A_{ij}(t - 1))$ if $t > 0$, in which $\hat{A}_{ij}(0) = A_{ij}(0)$ and we have the variational distributions

$$\rho_{ij}^{(1)}(t) = \frac{\lambda_{ij}}{\lambda_{ij} + \eta A_{ji}(t - 1)} \tag{14}$$

$$\rho_{ij}^{(2)}(t) = \frac{\eta A_{ji}(t - 1)}{\lambda_{ij} + \eta A_{ji}(t - 1)} \tag{15}$$

$$\phi_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k,q} u_{ik} v_{jq} w_{kq}}. \tag{16}$$

The parameter β has no closed-form update:

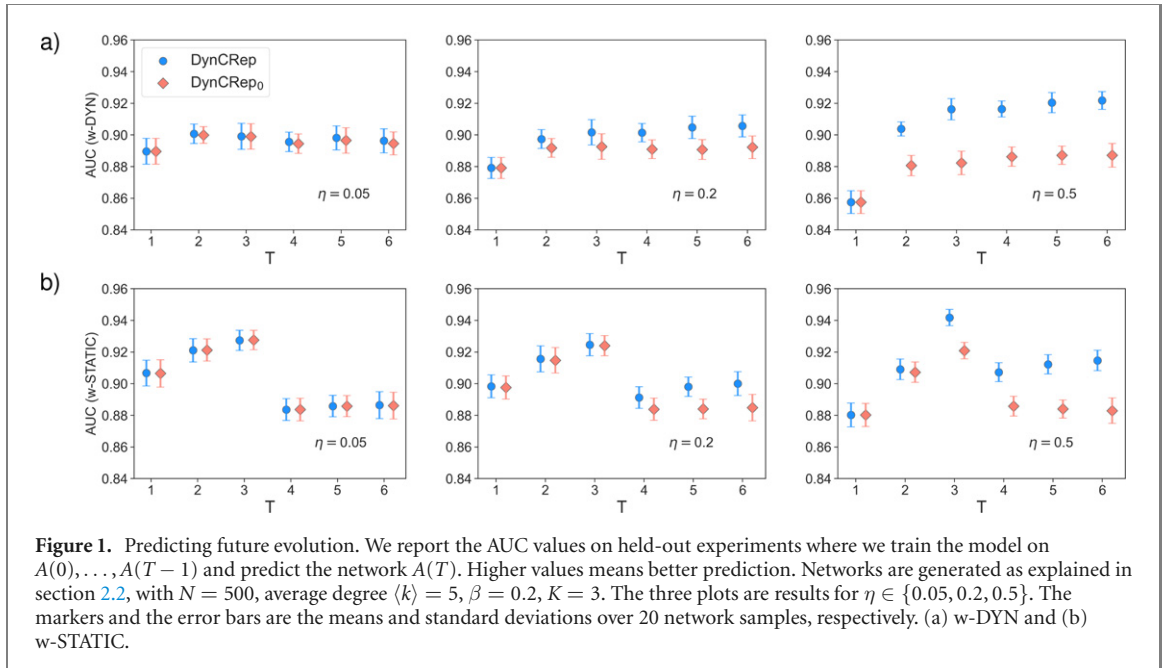
$$-\beta \left[T \sum_{i,j} \lambda_{ij} + \sum_{i,j,t=1}^{t=T} (\eta A_{ji}(t - 1)) + \frac{1}{1 - \beta} A_{ij}(t - 1) A_{ij}(t) \right] + \sum_{i,j,t=1}^{t=T} \left[\hat{A}(t) + A_{ij}(t - 1)(1 - A_{ij}(t)) \right] = 0, \tag{17}$$

but this equation can be solved numerically using root-finding methods. The algorithm proceeds by randomly initializing the parameters u, v, w, η, β ; then we estimate the variational distributions $\rho^{(1)}, \rho^{(2)}$, and ϕ , using equations (14)–(16) (E-step), while keeping the parameters fixed. In the next step (M-step), we update the parameters, while keeping $\rho^{(1)}, \rho^{(2)}$ and ϕ fixed. This procedure is repeated until the convergence of the likelihood in equation (9). An overview of the algorithm is described in algorithm 1.

2.2. Applications

2.2.1. Synthetic networks: AUC

Having explained the nuts and bolts of our model, we now turn to its application on dynamic network data. We start by considering synthetic networks generated by section 2.1 with known community structure and reciprocity. We assess the ability of the model in predicting the network at future time steps using past observations. We look in particular at the impact of reciprocity in determining edges, by generating networks with varying $\eta \in \{0.05, 0.2, 0.5\}$, while keeping the other parameters fixed.



For the tests reported here we use $N = 500$, initial average degree $\langle k \rangle = 5$, and $\beta = 0.2$. We generate $K = 3$ hard communities of equal size with assortative structure. Having fixed the parameters, we generate 20 samples of networks for each of the three values of η . For each network we generate an initial state followed by up to $T = 6$ further snapshots. The initial state is generated using only the community structure (no reciprocity) using equation (3). The successive snapshots are generated according to the instructions of section 2.1. In this study, to test the ability of our model in capturing the dynamical features, we generate the first three time snapshots ($T = 1, 2, 3$) with an assortative community structure and the rest of the snapshots ($T = 4, 5, 6$) with a disassortative community structure.

For each time step $t \in [1, T]$, we hide the individual snapshot $A(t)$ and fit the data using the previous snapshots $A(0), \dots, A(t-1)$. We test whether a model that accounts for reciprocity is able to successfully predict the network's evolution. Success is measured using the area under the curve (AUC), i.e., the probability that a randomly selected edge has higher expected value than a randomly selected non-existing edge. A value of 1 means perfect reconstruction, while 0.5 is pure random chance. The expected value of an edge is computed using:

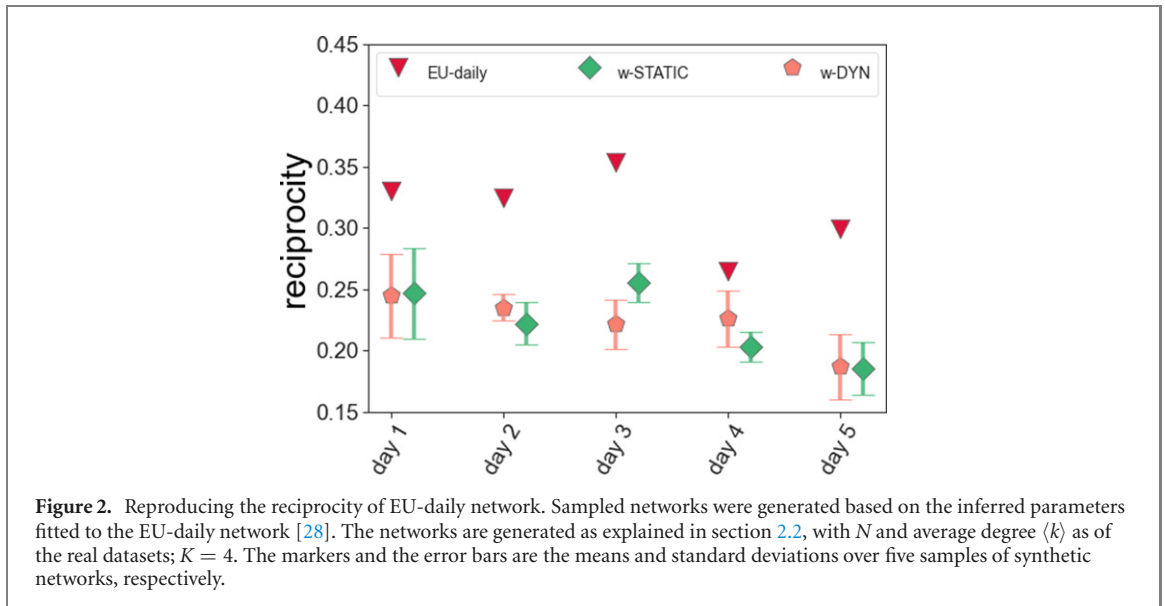
$$\mathbb{E}[A_{ij}(t)] = \begin{cases} \frac{p_{0 \rightarrow 1}}{p_{0 \rightarrow 1} + p_{0 \rightarrow 0}} & \text{if } A_{ij}(t-1) = 0 \\ \frac{p_{1 \rightarrow 1}}{p_{1 \rightarrow 1} + p_{1 \rightarrow 0}} & \text{if } A_{ij}(t-1) = 1 \end{cases} = \begin{cases} \beta(t)(\lambda_{ij}(t) + \eta A_{ji}(t-1)) & \text{if } A_{ij}(t-1) = 0 \\ 1 - \beta(t) & \text{if } A_{ij}(t-1) = 1. \end{cases} \quad (18)$$

Notice that while the expected value at time t uses explicitly only the network at the previous time step, all the parameters are inferred using the *whole* network history, i.e., the model is trained with $\{A(0), \dots, A(t-1)\}$. We compare with a model that does not account for reciprocity, i.e., our model with $\eta = 0$ (DynCRep₀) [25].

Figure 1 shows the results of these tests. As we can see, the ability to predict future edges is greater for a model that accounts for reciprocity, and the performance gap increases for higher values of η . This gap is partially offset by increasing the number of snapshots, as both the models have access to more information to make their estimates. Remarkably, DynCRep has stronger performance also in the low-reciprocity regime, $\eta = 0.05$. This cannot be clearly seen by looking at figure 1, as the mean AUC of the two models are within the error bars due to random fluctuations of the network structure across samples. Instead, the stronger performance of DynCRep in the low-reciprocity regime is revealed by looking at the percentage of samples where DynCRep has higher AUC than DynCRep₀, on a trial-by-trial case (see table 1 for details). While w-STATIC, the static version of the algorithm, performs slightly better than its non-reciprocated version, with larger performance gap at later times, w-DYN, the algorithm with time-varying affinity matrix, outperforms its non-reciprocated equivalent at all time steps.

Table 1. Edge prediction in synthetic networks. The stronger performance of DynCRep in the low-reciprocity regime, $\eta = 0.05$, is revealed by looking at the percentage of samples where DynCRep has higher AUC than DynCRep₀, on a trial-by-trial case, over 20 trials.

T	w-DYN		w-STATIC	
	DynCRep	DynCRep ₀	DynCRep	DynCRep ₀
1	0.0	0.0	57.0	43.0
2	71.0	29.0	43.0	57.0
3	86.0	14.0	38.0	62.0
4	67.0	33.0	43.0	57.0
5	71.0	29.0	52.0	48.0
6	81.0	19.0	57.0	43.0



Although both variants of the algorithm give better performance than their non-reciprocated version, it could be seen from figure 1 that w-DYN is more robust in link prediction tasks as η increases, and as the planted evolving structure of the affinity matrix changes from assortative to disassortative over time ($T = 4, 5, 6$).

2.2.2. Real world data: reciprocity/AUC

To evaluate the capability of our proposed model in retrieving network features, we apply the model to real world datasets. In this case, we first apply the inference algorithm to each time snapshot of the dynamic real dataset and learn the network's latent variables, i.e., Θ . Then, we use these latent variables as the input for the generative model, section 2.2, to generate dynamic synthetic networks similar to the fitted real datasets. Thus, we can compare dynamic synthetic networks, here 5 samples, and the original network. In this paper, we study the performance of our model in reproducing reciprocity as a significant structural parameter of the network. We implement our algorithm on two social and communication datasets, namely, email Eu core network [28] and statistics citation networks [29] (see section S6C for details on data pre-processing).

EU email network

Email-Eu-core network (EU) is constructed from internal emails exchanged between members of a large European research institution. At each time step, there is a directed edge from i to j , if i sent an email to j . Reciprocity may play a role in that receiving incoming emails may, or not, trigger a response email, similarly to other types of social communication [30]. The recorded dataset spans over a period of 803 days. However, we studied the dynamics of the dataset by dividing it in both daily and monthly durations. In the first case, we divide the edges in daily intervals (EU-daily); then select the snapshots from 5 consecutive days, randomly. In the latter case, the intervals are monthly; we select the snapshots from the first recorded year (EU-monthly).

Figure 2 shows the performance of w-DYN and w-STATIC versions of DynCRep in reproducing the reciprocity of the EU-daily network. As expected in email networks, the reciprocity is high in this case; hence, w-DYN and w-STATIC perform similarly in reproducing reciprocity. It is noticeable that the ability of reproducing reciprocity may change depending on how the network is built. For instance, if we consider the monthly time steps, EU-monthly network, we observe a different performance, see appendix S6B.

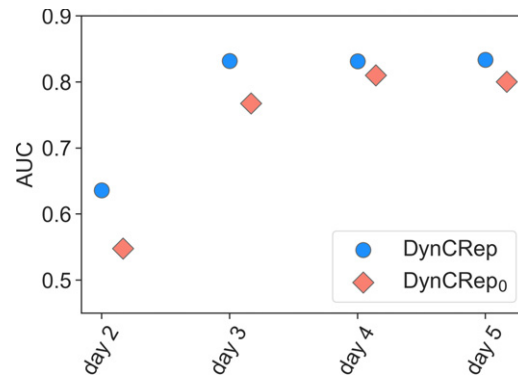


Figure 3. Predicting future evolution in the EU-daily dataset. AUC results for EU-daily dataset for five consecutive days, selected randomly. The number of community is fixed to $K = 4$. The error bars are smaller than marker size.

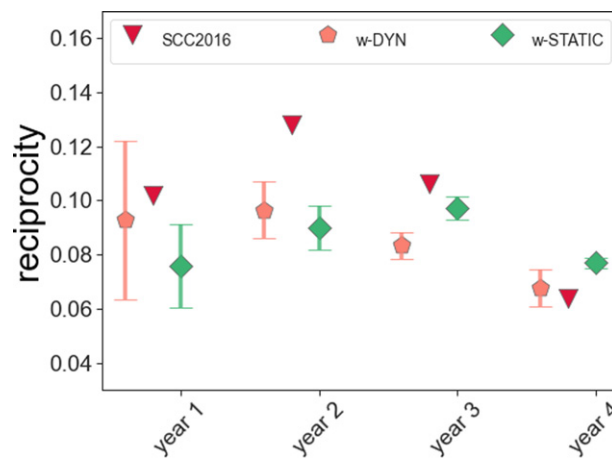


Figure 4. Reproducing the reciprocity of the statistics citation dataset. Sampled networks were generated based on the inferred parameters of the statistics citation dataset [29]. The networks are generated as explained in section 2.2, with N and average degree $\langle k \rangle$ as of the real datasets; $K = 3$. Markers and bars are the means and standard deviations over five generated synthetic networks, respectively. The network is based on annual citations during four years, from 2010 to 2013.

Figure 3 indicates the captured AUCs, measuring performance in link prediction tasks. The AUC is calculated as described in section 2.2.1. We can notice the improvement over the time snapshots, and DynCRep tends to perform slightly better. Therefore by having access to the history of the dataset and accounting for reciprocity we can achieve better results in predicting future connections.

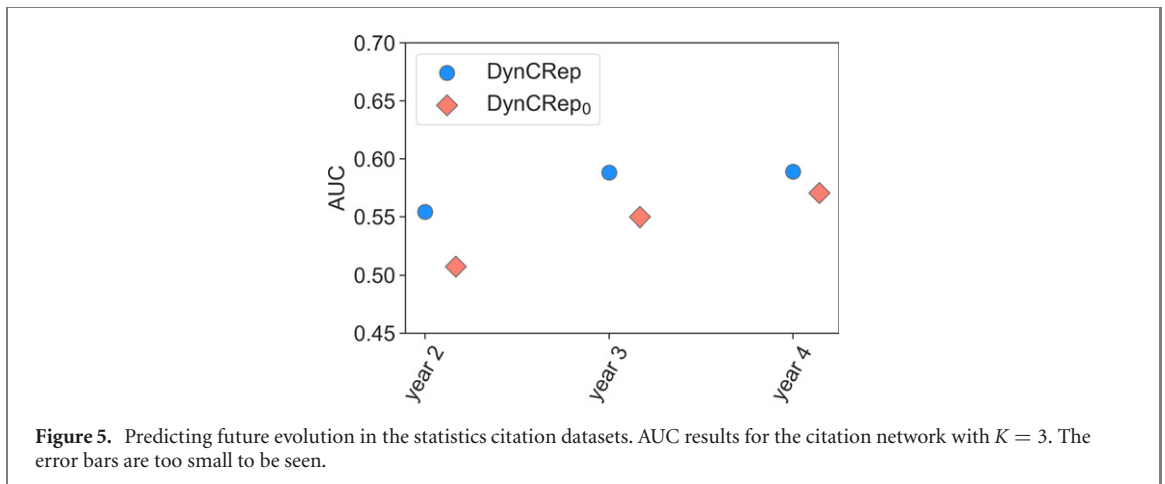
It is worth mentioning that we performed the experiments for different values of the number of communities; however, the results do not show high sensitivity to this parameter. Therefore, we fixed $K = 4$ for the EU network, equivalent to the number of departments in the corresponding institute.

Statistics citation dataset

The second example of an empirical dataset is the citation networks for statisticians, which is based on the research papers published in four of the top journals in statistics from 2003 to the first half of 2012. We construct a network by selecting a sample of the data from 2003 to 2007 and dividing it into annual intervals. This way we will have a network of citations over 4 years, where nodes are authors and an edge from nodes i to j at time step T represents that i cites j 's papers in that year. In this system, we may expect that reciprocity plays a role in that receiving a citation may trigger a citation back.

Despite the fact that the reciprocity in this dataset is much lower than EU-daily dataset, figure 4 shows that we are able to capture it competitively. In addition, although the two versions outperform each others at different time steps, they still behave similarly in reproducing the reciprocity. Moreover, in both empirical datasets, the best performance is obtained for the case that reciprocated edges presented at the same time step were used in the model.

As it could be seen from figure 5, AUC values are always higher for DynCRep, showing that accounting for reciprocity improves link prediction tasks also for this dataset. It should be noted that, at each time step T we calculate AUC by having access to the edges up to time $T - 1$, then predicting edges at time T . Hence, the AUC



cannot be calculated for the first time step. In this case we fix $K = 3$, the minimum number of communities with the highest performance, i.e., we perform five-fold cross validation [25] to calculate the value of AUC, then we choose K as the number of communities with the highest value for AUC.

3. Conclusion

In this work, we study reciprocity in dynamic networks. In reality, many datasets, e.g., networks of friendship, of gene expression patterns or communication networks, describe interactions that evolve over time, thus making them unsuitable objects of analysis for aggregate methods. In addition, the interactions in these networks might not simply change over time, but their evolution could also be affected by their past reciprocated interactions; generally, such reciprocal interactions have received little attention as additional drivers of this dynamics.

To remedy this problem, we combine insights from previous works to incorporate reciprocity into a generative model approach with latent community structure. Specifically, we extend the assumptions formulated in [25] to situations where networks change in time. For this, we consider a Markovian transition matrix which governs the evolution of the parameters over time snapshots. Being a generative model, our approach can be used to build dynamic synthetic networks, with desired reciprocity and community structure. Its algorithmic implementation is based on an efficient EM algorithm, which can be applied to large systems. As we assume a chronological order in observing the reciprocated edges, we can estimate the joint probability distribution as a factorized distribution of time steps.

We consider two varieties of our model. In one case, community membership vectors remain static over time and only the affinity matrix contains temporal information. In the other case, the affinity matrix is treated as a static parameter, similarly as the community memberships; in both cases, reciprocity parameter and the rate of edge removal are kept static. These two scenarios enable us to thoroughly analyze the model and its performance in networks with different interaction patterns. For instance, in the case of a non-homogeneous community structure over time, the first version would be a more suitable approach, since it could capture the evolving community structures.

There are a number of directions in which this model could be extended. To capture more realistic properties of the real world datasets, we can generalize the model to the case of multilayered networks, where nodes can have more than one type of interaction. For instance, in a social network, an individual can have connections based on friendship, as well as her business affiliations.

In addition, considering a node related reciprocity parameter instead of a global reciprocity parameter could improve the applicability of the model. We have focused here on the case where edges change in time, but one can envisage situations where nodes appear and disappear as well. This would also be a natural model extension. Finally, we considered here reciprocity as main network structural property, but similar investigations can be performed for other properties involving more than one pair of nodes, as triadic closure or transitivity.

Acknowledgments

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani. All the authors were supported by the Cyber Valley Research Fund.

Data and code availability

Any data that support the findings of this study are included within the article. We will provide an open source version of the code online upon publication.

ORCID iDs




Hadiseh Safdari  <https://orcid.org/0000-0002-3814-2640>

Martina Contisciani  <https://orcid.org/0000-0002-6103-5499>

Caterina De Bacco  <https://orcid.org/0000-0002-8634-0211>

References

- [1] Vespignani A 2012 *Nat. Phys.* **8** 32
- [2] Holme P 2015 *Eur. Phys. J. B* **88** 234
- [3] Scholtes I 2017 *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '17* (New York: ACM) pp 1037–46
- [4] Mandjes M, Starrevelde N, Bekker R and Spreij P 2019 Dynamic Erdős–Rényi graphs *Computing and Software Science: State of the Art and Perspectives* ed B Steffen and G Woeginger (Berlin: Springer) pp 123–40
- [5] Holland P W, Laskey K B and Leinhardt S 1983 *Soc. Network.* **5** 109
- [6] Wang Y J and Wong G Y 1987 *J. Am. Stat. Assoc.* **82** 8
- [7] Nowicki K and Snijders T A B 2001 *J. Am. Stat. Assoc.* **96** 1077
- [8] Karrer B and Newman M E J 2011 *Phys. Rev. E* **83** 016107
- [9] De Bacco C, Power E A, Larremore D B and Moore C 2017 *Phys. Rev. E* **95** 042317
- [10] Yang T, Chi Y, Zhu S, Gong Y and Jin R 2011 *Mach. Learn.* **82** 157
- [11] Matias C and Miele V 2017 *J. R. Stat. Soc. B* **79** 1119
- [12] Zhang X, Moore C and Newman M E 2017 *Eur. Phys. J. B* **90** 200
- [13] Xu K S and Hero A O 2014 *IEEE J. Sel. Top. Signal Process.* **8** 552
- [14] Han Q, Xu K S and Airolidi E M 2014 arXiv:1410.8597
- [15] Peixoto T P and Rosvall M 2017 *Nat. Commun.* **8** 582
- [16] Matias C, Rebafka T and Villers F 2018 *Biometrika* **105** 665
- [17] Gauvin L, Panisson A and Cattuto C 2014 *PLoS One* **9** e86028
- [18] Bovet A, Delvenne J-C and Lambiotte R 2021 arXiv:2101.06131
- [19] Rossetti G and Cazabet R 2018 *ACM Comput. Surv.* **51** 1
- [20] Ghasemian A, Zhang P, Clauset A, Moore C and Peel L 2016 *Phys. Rev. X* **6** 031005
- [21] Mucha P J, Richardson T, Macon K, Porter M A and Onnela J-P 2010 *Science* **328** 876
- [22] Herlau T, Mørup M and Schmidt M 2013 *Int. Conf. on Machine Learning* (PMLR) pp 960–8
- [23] Holland P W and Leinhardt S 1981 *J. Am. Stat. Assoc.* **76** 33
- [24] Garlaschelli D and Loffredo M I 2004 *Phys. Rev. Lett.* **93** 268701
- [25] Safdari H, Contisciani M and De Bacco C 2021 *Phys. Rev. Res.* **3** 023209
- [26] Bartolucci F, Marino M F and Pandolfi S 2018 *Comput. Stat. Data Anal.* **123** 86
- [27] Gardiner C W 2004 *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer Series in Synergetics) 3rd edn vol 13 (Berlin: Springer) pp xviii+415
- [28] Leskovec J, Kleinberg J and Faloutsos C 2007 *ACM Trans. Knowl. Discov. Data* **1** 2
- [29] Ji P and Jin J 2016 *Ann. Appl. Stat.* **10** 1779
- [30] Aoki T, Takaguchi T, Kobayashi R and Lambiotte R 2016 *Phys. Rev. E* **94** 042313

Anomaly, reciprocity, and community detection in networksHadiseh Safdari ^{*}, Martina Contisciani [†], and Caterina De Bacco [‡]
Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen 72076, Germany

(Received 2 February 2023; accepted 22 July 2023; published 7 August 2023)

Anomaly detection algorithms are a valuable tool in network science for identifying unusual patterns in a network. These algorithms have numerous practical applications, including detecting fraud, identifying network security threats, and uncovering significant interactions within a data set. In this project, we propose a probabilistic generative approach that incorporates community membership and reciprocity as key factors driving regular behavior in a network, which can be used to identify potential anomalies that deviate from expected patterns. We model pairs of edges in a network with exact two-edge joint distributions. As a result, our approach captures the exact relationship between pairs of edges and provides a more comprehensive view of social networks. Additionally, our study highlights the role of reciprocity in network analysis and can inform the design of future models and algorithms. We also develop an efficient algorithmic implementation that takes advantage of the sparsity of the network.

DOI: [10.1103/PhysRevResearch.5.033084](https://doi.org/10.1103/PhysRevResearch.5.033084)**I. INTRODUCTION**

Anomaly detection algorithms are a crucial tool in the study of networks. These algorithms are designed to identify unusual or unexpected patterns in the data, which can provide valuable insights into the structure and function of a network [1,2]. For instance, anomalous edges in a network may indicate the presence of a structural flaw or a potential problem, such as a vulnerability to attack. By detecting and analyzing these anomalies, we can gain a better understanding of the network and potentially identify ways to improve its performance or security [3]. In addition, anomaly detection algorithms can be used to monitor networks in real time, allowing researchers to quickly identify and respond to potential issues as they arise.

Anomalies are often difficult to define precisely because they can vary depending on the context and the system being analyzed [4]. For example, in a network of online transactions, an anomaly could be a sudden spike in the number of transactions coming from a single user [5]. In this case, the regular behavior in the system would be the typical number of transactions coming from a single user, and any deviation from this pattern would be considered an anomaly. Hence, one of the main obstacles in detecting anomalies in networks is determining what is considered “normal” (or “regular”)

behavior. To overcome this challenge, we must create a null model which is a realistic representation of the network data. This null model provides a standard against which we can compare the network data and identify anomalies.

Relevant approaches to address this problem include statistics-based methods, which fit a statistical model to the network data [6,7]. Among these, generative models [8–10] make assumptions about the processes that drive network formation and evolution to generate synthetic network data. By using these approaches, we can define null models that are tailored to the specific characteristics of the network under study. This is the approach we take here.

In this work, we focus on plain networks, which only contain information about the presence or absence of connections between individuals, and do not include any additional information. One approach to perform anomaly detection in these binary and single-layer networks is to use the structure of the graph to identify patterns and detect deviations from them [1]. These structural patterns can be divided into two categories: Patterns based on the *overall structure* of the graph, and patterns based on the *community structure* of the graph. Methods in the first category rely on the global properties of the graph [11], such as the distribution of node degrees or the overall connectivity of the network. On the other hand, methods in the second category perform anomaly detection by focusing on the local properties of the graph, such as the membership of nodes in communities [12,13]. Hence, with the second approach, we assume that the null model reflects a community structure that can be identified through latent variables, a process known as community detection task [14]. Thus, by considering the community structure, anomalous behavior can be determined in this context. For example, a friendship between two individuals from different groups, such as high school classmates and college classmates, could be considered anomalous. We recently developed a model anomaly community detection (ACD) that performs anomaly detection by

^{*}hadiseh.safdari@tuebingen.mpg.de[†]martina.contisciani@tuebingen.mpg.de[‡]caterina.debacco@tuebingen.mpg.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by the Max Planck Society.

using community structure [15], where anomalous edges are those that deviate from regular patterns determined by community structure. As a result, this model outputs both node memberships and edge labels, identifying them as legitimate or anomalous.

Several notable studies have tackled the inherent challenges associated with reconstructing networks from unreliable and imperfect measurements [16–18], a problem related to the subject of this study. These methods generate a posterior distribution encompassing potential network structures in the presence of errors and uncertainties. Consequently, these approaches provide probabilities indicating the presence or absence of edges, whereby lower probabilities can be interpreted as indicative of less reliable connections. For instance, De Bacco *et al.* [16] specifically address the task of handling noisy and multiply reported social network data by introducing latent network models capable of accommodating errors and uncertainties in the observed data. By incorporating latent variables, this model successfully captures the underlying true network structure while accounting for reporting errors. Similarly, Peixoto [17] presents a framework that enhances reconstruction accuracy by employing generative network methods assuming modular structures. Furthermore, Newman [18] proposes a general probabilistic model to estimate network structure from unreliable network data. Although these studies may provide some indication of tie unreliability, their primary focus lies in network reconstruction rather than in anomaly detection, as we do here. Therefore, they do not incorporate an explicit structure responsible for the presence of anomalies which allows to thoroughly investigate their nature.

In fact, accurately identifying anomalies is deeply connected with the chosen null model determining what regular patterns are. As a consequence, it is important to consider other possible mechanisms for tie formation, beyond community structure. For instance, reciprocity, another fundamental structural feature in networks [16,19,20], refers to the mutual exchange of resources or actions between individuals or groups. This can include actions such as returning a favor, sharing information or resources, or collaborating on a project. For example, in a social network, if two individuals consistently like and comment on each other's posts, this could be considered reciprocity. In a business network, if two companies frequently refer customers to each other, this could also be considered reciprocity. In both cases, reciprocity implies a two-way exchange rather than a one-sided action. A reciprocated edge refers to a connection between two nodes in the network where both nodes have engaged in mutual interactions. Mathematically, reciprocity is calculated as the ratio of the number of reciprocated edges to the total number of edges in the graph. A reciprocated edge is formed when there is a bidirectional relationship between the connected nodes, signifying that both nodes have reciprocated actions or engagements.

Recent works [21,22] have shown that including reciprocity effects in the modeling of community patterns results in more accurate and expressive generative models. This has the potential to improve the performance of an anomaly detection model for networks as well.

In this work, we develop a probabilistic generative model that we refer to as the community reciprocity anomaly detection (CRAD) algorithm, which performs anomaly detection by proposing a null model based on both community structure and reciprocity. Intuitively, our model regards as regular ties those who follow the group membership and reciprocity effects, and as anomalous ties those whose formation process is not aligned with these two mechanisms. Notice that node memberships, reciprocity effect, and anomalous edges are all unknown processes. Our model is able to infer them from data by representing them as latent variables in a probabilistic model.

More specifically, we model the existence of ties between pairs of nodes using a bivariate Bernoulli distribution. This has the crucial statistical property that independence and uncorrelatedness of the component random variables are equivalent [23], which facilitates the derivation of a closed-form joint distribution of a pair of edges. Furthermore, both the marginal and conditional distributions are Bernoulli distributions, enabling closed-form analytical expressions. This facilitates downstream analysis and also improves model performance, as shown in Ref. [21].

II. THE MODEL

We are given an adjacency matrix, \mathbf{A} , as our observed data, with entries indicating the presence or absence of an edge from node i to node j , represented by $A_{ij} = 1$ or $A_{ij} = 0$, respectively. Pairs of directed edges between two nodes (i, j) are defined as $A_{(ij)} = (A_{ij}, A_{ji})$. We consider binary data, thus $A_{(ij)} \in \{0, 1\}^2 = \{0, 1\} \times \{0, 1\}$, and directed networks, i.e., in general $A_{ij} \neq A_{ji}$. We aim at classifying any such pair as either regular or anomalous, accounting for community structure and reciprocity effects. For this, we introduce a Bernoulli random variable that represents the binary label of being anomalous or not as a random variable:

$$\sigma_{(ij)} \sim \text{Bern}(\mu), \quad (1)$$

where $\sigma_{(ij)} = 0, 1$ if the pair $A_{(ij)}$ is regular or anomalous, respectively. In this work we assume that edges between any pair of nodes must be either anomalous or regular. Mathematically, this means that the matrix σ with entries σ_{ij} is symmetric, i.e., $\sigma_{ij} = \sigma_{ji}$. These latent variables must be learned from data, as anomalies are not known in advance. They also determine the mechanism from which the pair of edges is drawn. The hyperparameter $\mu \in [0, 1]$ controls the prior distribution of $\sigma_{(ij)}$.

With these main ingredients in mind, we can proceed to characterize the joint probability distribution of pairs of edges. Assuming to know the label $\sigma_{(ij)}$ for a given pair of edges, we denote the pair joint probability $p_{nm}^{(\ell)} = P^{(\ell)}(A_{ij} = n, A_{ji} = m)$, where $n, m \in \{0, 1\}$ and $\ell \in \{r, a\}$ denotes the label being regular or anomalous, respectively. We then consider the joint probability distribution of a pair of edges as a bivariate

Bernoulli distribution:

$$\begin{aligned}
P(A_{(ij)}, \sigma_{(ij)}) &= P(A_{ij}, A_{ji}, \sigma_{(ij)}) = P(A_{ij}, A_{ji} | \sigma_{(ij)}) P(\sigma_{(ij)}) \\
&= P^{(a)}(A_{ij}, A_{ji} | \theta_a)^{\sigma_{(ij)}} P^{(r)}(A_{ij}, A_{ji} | \theta_r)^{1-\sigma_{(ij)}} P(\sigma_{(ij)} | \mu) \\
&= \left[[P_{11}^{(a)}]^{A_{ij} A_{ji}} [P_{10}^{(a)}]^{A_{ij}(1-A_{ji})} [P_{01}^{(a)}]^{(1-A_{ij})A_{ji}} [P_{00}^{(a)}]^{(1-A_{ij})(1-A_{ji})} \right]^{\sigma_{(ij)}} \\
&\quad \times \left[[P_{11}^{(r)}]^{A_{ij} A_{ji}} [P_{10}^{(r)}]^{A_{ij}(1-A_{ji})} [P_{01}^{(r)}]^{(1-A_{ij})A_{ji}} [P_{00}^{(r)}]^{(1-A_{ij})(1-A_{ji})} \right]^{1-\sigma_{(ij)}} \mu^{\sigma_{(ij)}} (1-\mu)^{1-\sigma_{(ij)}}, \quad (2)
\end{aligned}$$

where θ_r and θ_a denote parameters specific to the two distributions $P^{(r)}$ and $P^{(a)}$. The parameters $p_{nm}^{(\ell)}$ must satisfy $\sum_{n,m=0,1} p_{nm}^{(\ell)} = 1$ to have valid probability density functions.

Following the notation as in Refs. [21,23], we can rewrite the full joint probability density function in Eq. (2) as the product

$$P(\mathbf{A}, \boldsymbol{\sigma}) = \prod_{(i,j)} \left[\frac{\exp \{A_{ij} f_{ij}^{(a)} + A_{ji} f_{ji}^{(a)} + A_{ij} A_{ji} J_{(ij)}^{(a)}\}}{Z_{(ij)}^{(a)}} \times \mu \right]^{\sigma_{(ij)}} \left[\frac{\exp \{A_{ij} f_{ij}^{(r)} + A_{ji} f_{ji}^{(r)} + A_{ij} A_{ji} J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} \times (1-\mu) \right]^{1-\sigma_{(ij)}}, \quad (3)$$

where $p_{00}^{(\ell)} = 1/Z_{(ij)}^{(\ell)}$, and $Z_{(ij)}^{(\ell)}$ is the normalization constant for the regular or anomalous edges, for $\ell \in \{r, a\}$; $f_{ij}^{(\ell)}$, $f_{ji}^{(\ell)}$, and $J_{(ij)}^{(\ell)}$ are the natural parameters of their density functions. The interaction term $J_{(ij)}^{(\ell)}$ appears in order to capture reciprocity. It allows to have a joint pair distribution $P(A_{ij}, A_{ji} | \sigma_{(ij)})$ that is not simply the product of two independent distributions $P(A_{ij} | \sigma_{(ij)}) \times P(A_{ji} | \sigma_{(ij)})$, as it is usually assumed in cases where reciprocity (or other properties involving more than one variable) is not taken into account explicitly.

These parameters can be expressed in terms of the probability of occurrence of edges as follows:

$$\begin{aligned}
f_{ij}^{(\ell)} &= \log \left(\frac{p_{10}^{(\ell)}}{p_{00}^{(\ell)}} \right), \quad f_{ji}^{(\ell)} = \log \left(\frac{p_{01}^{(\ell)}}{p_{00}^{(\ell)}} \right), \\
J_{(ij)}^{(\ell)} &= \log \left(\frac{p_{11}^{(\ell)} p_{00}^{(\ell)}}{p_{10}^{(\ell)} p_{01}^{(\ell)}} \right), \quad \ell = \{r, a\}. \quad (4)
\end{aligned}$$

We aim at modeling reciprocity when two edges are regular, as this can be the result of a reasonable tie formation mechanism involving two nodes, e.g., exchanging favors or cooperative behaviors. For anomalous edges instead, it is less clear what would reciprocity mean; hence we remain agnostic to it and assume that the edges $i \rightarrow j$ and $j \rightarrow i$ are independent when they are anomalous. In other words, the existence of the anomalous edge A_{ji} has no influence on its reciprocated edge A_{ij} , which is also anomalous. To reflect this mathematically, we set $J_{(ij)}^{(a)} = 0$. This follows the properties of multivariate Bernoulli distributions, where independence and uncorrelatedness are equivalent phenomena [23]. As the correlation between the pair of edges (A_{ij}, A_{ji}) is captured by $J_{(ij)}^{(\ell)}$, when $J_{(ij)}^{(\ell)} = 0$, the pair (A_{ij}, A_{ji}) is uncorrelated. In addition, we assume a symmetric structure of $f^{(a)} = f_{ij}^{(a)} = f_{ji}^{(a)}$ for all anomalous edges.

To summarize the steps of our proposed generative model: We first draw hidden labels for the edges, determining them to be regular or anomalous; then, we draw pairs of edges (A_{ij}, A_{ji}) from a specific form of distribution depending on the edges' labels. Formally, the generative model is

$$\sigma_{(ij)} \sim \text{Bern}(\mu), \quad (5)$$

$$A_{(ij)} \sim \begin{cases} \frac{\exp \{(A_{ij} + A_{ji}) f^{(a)}\}}{Z_{(ij)}^{(a)}} & \text{if } \sigma_{(ij)} = 1 \\ \frac{\exp \{A_{ij} f_{ij}^{(r)} + A_{ji} f_{ji}^{(r)} + A_{ij} A_{ji} J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} & \text{if } \sigma_{(ij)} = 0. \end{cases} \quad (6)$$

Up to this point, we focused on reciprocity and how to incorporate it into our model via the interaction term $J_{(ij)}^{(r)}$. Now, we turn our attention to community structure, another important mechanism that we believe regulates tie formation of regular edges. Conversely, we assume that communities have no influence on anomalous edges. To formalize this, we utilize similar model specifications as outlined in Ref. [21], and we incorporate community structure through latent variables embedded in the natural parameters of the joint pair distribution $P^{(r)}(A_{ij}, A_{ji} | \theta_r)$. In detail, we assume the tie formation depends on communities and reciprocity for regular edges, and only on the anomaly parameter for anomalous ties:

$$f_{ij}^{(r)} = \log \lambda_{ij}, \quad f_{ji}^{(r)} = \log \lambda_{ji}, \quad (7)$$

$$J_{(ij)}^{(r)} = \log \eta, \quad (8)$$

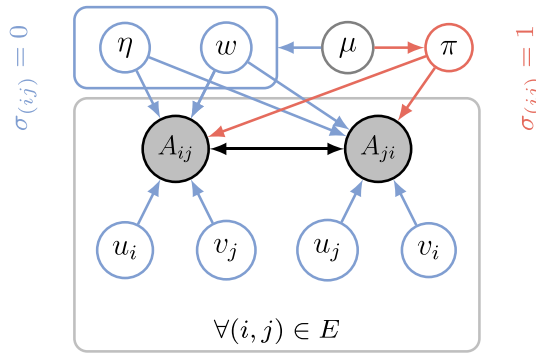
$$f^{(a)} = \log \pi, \quad (9)$$

where

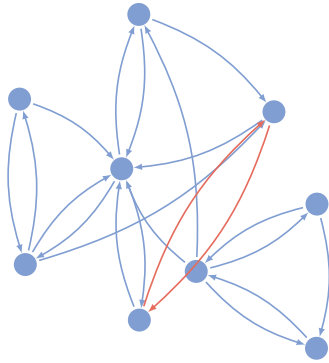
$$\lambda_{ij} = \sum_{k,q=1}^K u_{ik} v_{jq} w_{kq} \quad (10)$$

regulates how mixed-membership community structure determines tie formation in directed networks, as in Ref. [24]. We provide a schematic visualization of these contributions in Fig. 1. The normalization parameters are obtained by enforcing the normalization constraint using the above definitions, so that $Z_{(ij)}^{(a)} = (\pi + 1)^2$ and $Z_{(ij)}^{(r)} = \lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1$.

The parameters λ and η play important roles in our model of community-reciprocity structure. λ captures the mixed-membership aspect, while η is the pair-interaction coefficient that regulates the formation of pairs of edges between nodes. The K -dimensional vectors u_i and v_i represent the outgoing and incoming communities of node i , respectively. The entries in these vectors, $u_{ik} \geq 0$ and $v_{jq} \geq 0$, represent the weights assigned to each community, where K is the number of communities. The value of K can be either specified



(a) Graphical model representation.



(b) Network example.

FIG. 1. Model visualization. (a) Graphical model: The entry of the adjacency matrix A_{ij} is determined by the community-related latent variables u, v , and w and the reciprocity parameter η (blue), and by the anomaly-related parameters π (orange) and the hyperprior μ (grey). (b) Example of a possible realization of the model: Blue edges display interactions based on community and reciprocity and the orange ones are anomalous.

as input or selected using model selection criteria, such as cross-validation [24]. The affinity matrix w_{kq} controls the structure of the communities, with higher values on the diagonal indicating more assortative communities. The formation of anomalous edges is derived by the latent parameter $\pi > 0$, as in the $\lim \pi \rightarrow 0$ the probability of the existence of an anomalous edge converges to zero (see Appendix A for more details on derivations). All of these parameters, along with μ , are included in the latent parameter set $\Theta = \{\{u_i\}, \{v_i\}, \{w_{kq}\}, \eta, \pi, \mu\}$ that will be inferred from data. In addition to point estimates of these parameters, our model returns a posterior estimate for the edge label variable $\sigma_{(ij)}$ in the form of a Bernoulli posterior distribution of parameter $Q_{(ij)}$. This is also the estimated expected value of the edge label. We provide more details in Sec. III.

Our model assumes that community structure drives the process of formation of a regular edge, and that the regular edges between a pair of nodes depend on each other explicitly according to the value of η . If $J_{(ij)}^{(r)} = 0$ (when $\eta = 1$), the probability of the edges between nodes i and j is determined solely by their respective communities. On the other hand, a positive value of $J_{(ij)}^{(r)}$ (when $\eta > 1$) increases the probability

of the existence of both $i \rightarrow j$ and $j \rightarrow i$, while a negative value (when $0 < \eta < 1$) decreases it.

By utilizing properties of the bivariate Bernoulli distribution [21,23], we obtain a closed-form solution for the expected value of an edge (see Appendix A for more details):

$$\mathbb{E}[A_{ij}] = (1 - Q_{(ij)}) \frac{\lambda_{ij} + \eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}^{(r)}} + Q_{(ij)} \frac{\pi}{1 + \pi}. \quad (11)$$

This result is useful in link prediction experiments, in that we can score edges based on the values calculated from Eq. (11) and use these to compute prediction metrics such as the area under the receiver operating curve (AUC). We illustrate this in Sec. IV A.

III. INFERENCE

Our ultimate goal is to determine Θ , the latent parameters of the model. To do this, we maximize the posterior probability $P(\Theta|A) = \sum_{\sigma} P(\sigma, \Theta|A)$. Instead of directly maximizing this probability, it is more computationally efficient to maximize the log-posterior, as the maxima of the two functions are equivalent:

$$\begin{aligned} L(\Theta) &= \log P(\Theta|A) = \log \sum_{\sigma} P(\sigma, \Theta|A) \\ &\geq \sum_{\sigma} q(\sigma) \log \frac{P(\sigma, \Theta|A)}{q(\sigma)}, \end{aligned} \quad (12)$$

where we defined $q(\sigma)$, a variational distribution that must sum to 1. Our maximum likelihood approach involves the use of an expectation-maximization (EM) algorithm in which we alternately update different sets of parameters of our model. More specifically, we first update the variational distribution parameters (E step), ρ and Q , and then maximize $L(\Theta)$ with respect to Θ (M step). This process is repeated until $L(\Theta)$ converges, signifying the completion of the optimization process. The full procedure is outlined in Algorithm 1 (see Appendix B for more details on the inference task). The computational complexity of the algorithm is $O(N^2)$, primarily due to the terms in the dense matrix $Q_{(ij)}$ that are not multiplied by the sparse adjacency matrix A_{ij} . As Q is crucial for identifying anomalous edges, its presence may make the model infeasible for large systems. Investigating ways to reduce this complexity, for instance by making its representation sparse, is an interesting avenue for future work.

IV. RESULTS

A. Synthetic data sets

We validate our model on synthetic data sets, generated with the generative algorithm in Appendix C. The studied data sets consist of $N = 500$ nodes, with an average degree of $\langle k \rangle = 60$. The number of communities is set to $K = 3$, and the pair-interaction coefficient, η , has a range of values. The anomaly density (ratio of anomalous edges to total number of edges) is varied within the interval $\rho_a \in [0, 1]$. We compare CRAD with JointCRep [21], which is what CRAD reduces to if we had not considered anomalies, i.e., when $\mu = 0$ and $\lim \pi \rightarrow 0$. This allows to focus on observing the impact of

Algorithm 1: CRAD: EM algorithm.

Input: network $A = \{A_{ij}\}_{i,j=1}^N$, number of communities K .

Output: memberships $u = [u_{ik}]$, $v = [v_{ik}]$; network affinity matrix $w = [w_{kq}]$; pair-interaction coefficient η ; anomaly parameter π ; prior on anomaly indicator μ .
Initialize $\Theta : (u, v, w, \eta, \pi, \mu)$ at random.Repeat until $L(\Theta)$ convergence:1. Calculate ρ and Q (E step):

$$\rho_{ijkq} \sim \text{as in Eq. (B13)},$$

$$Q_{ij} \sim \text{as in Eq. (B22)}.$$

2. Update parameters Θ (M step):(i) For each node i and community k update memberships:

$$u_{ik} = \frac{\sum_{jq} (1 - Q_{ij}) A_{ij} \rho_{ijkq}}{\sum_j \left[\frac{\sum_q (1 - Q_{ij}) (1 + \eta \lambda_{ji}) v_{jq} w_{kq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]},$$

$$v_{ik} = \frac{\sum_{jq} (1 - Q_{ij}) A_{ji} \rho_{jikq}}{\sum_j \left[\frac{\sum_q (1 - Q_{ij}) (1 + \eta \lambda_{ij}) u_{jq} w_{kq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}.$$

(ii) For each pair (k, q) update affinity matrix:

$$w_{kq} = \frac{\sum_{i,j} (1 - Q_{ij}) A_{ij} \rho_{ijkq}}{\sum_{i,j} \left[\frac{(1 - Q_{ij}) (1 + \eta \lambda_{ij}) u_{ik} v_{jq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}.$$

(iii) Update pair-interaction coefficient:

$$\eta = \frac{\sum_{(i,j)} (1 - Q_{ij}) A_{ij} A_{ji}}{\sum_{(i,j)} (1 - Q_{ij}) \left[\frac{\lambda_{ij} \lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}.$$

(iv) Update anomaly parameter:

$$\pi = \frac{\sum_{(i,j)} Q_{ij} (A_{ij} + A_{ji})}{\sum_{(i,j)} Q_{ij} (2 - A_{ij} - A_{ji})}.$$

(v) Update prior on anomaly indicator:

$$\mu = \frac{1}{N(N-1)/2} \sum_{(i,j)} Q_{ij}.$$

considering the existence of anomalous edges in a given data set.

In order to determine the effectiveness of our proposed model, which is based on the concept of community structure, we first evaluate its ability to accurately identify the memberships of individuals within a community. To accomplish this, we measure the cosine similarity (CS) between the ground truth and inferred community membership vectors. The CS has values in $[0,1]$, with $CS = 1$ indicating the best performance. For this task, we also run a Bayesian Poisson matrix

factorization (BPMF) algorithm [25]. BPMF is a scalable algorithm for factorizing sparse matrices and provides a useful comparison for our proposed algorithm. We run all algorithms on synthetic data sets generated by CRAD (see Appendix C for more details). The results, as illustrated in Figs. 6(a) and 6(b), show that when the proportion of anomalous edges in the data set is relatively low, BPMF outperforms our proposed algorithm. However, when the number of anomalous edges is above 50% of the total number of edges, our algorithm is still able to detect community structure with a reasonable level of accuracy. Additionally, it can be observed that CRAD performs the same as JointCRep, with both models having higher performance for smaller values of the anomaly density, ρ_a . This behavior is expected, as for higher values of ρ_a , the community structure plays a weaker role in the formation of edges.

It is worth mentioning that the primary objective of the current research is to develop the capabilities of JointCRep through the incorporation of anomaly detection functionality, rather than focusing on further improving its community detection abilities or recovering reciprocity parameter. Therefore, our focus is on assessing and optimizing the model's anomaly detection potential. For this, we measure the AUC on edges, i.e., on a binary matrix that stores what edges are true anomalies, and use as scores the inferred Q_{ij} . From our results, illustrated in Figs. 6(e) and 6(f), we find that CRAD demonstrates good performance in the detection of anomalous edges across a range of anomaly densities. Furthermore, the integration of reciprocity effects is enhancing performance, compared to a model (ACD) where there is no such effect [15].

In addition to evaluating anomaly detection, we are also interested in assessing the ability of CRAD to identify missing edges, also known as the link prediction task. In these experiments we employ a fivefold cross-validation approach, where the data set is split into five sets of data. In each realization, four of these groups are utilized as a training set to infer the parameters Θ . The remaining group is used as a test set, where the score for each pair (A_{ij}, A_{ji}) in the matrix is evaluated to compute the AUC. By iteratively varying which group serves as the test set, we obtain a total of five trials per realization. The final AUC value is determined by averaging the results of these trials. The score of an edge is calculated using the closed-form expression for its marginal probability, as described in Eq. (11). As shown in Figs. 6(c) and 6(d), an increase in the reciprocity parameter results in an increase in the AUC for both CRAD and JointCRep; however, we observe a bigger improvement in terms of AUC of CRAD over the competitive algorithms. These results indicate that our model becomes more effective in link prediction tasks for higher values of reciprocity.

B. Real-world data sets

In order to assess the practical utility of our model, we investigate its usage on a variety of real-world data covering applications such as food-sharing between bats, social support interactions in a rural community, email exchanges, and on-line dating. Their sizes range from $N = 19$ to $N = 3562$ (see Table I for a summary description). We select the number of

TABLE I. Real-world data-set descriptions.

Network	Abbreviation	N	E	$\langle k \rangle$	Reciprocity	Ref.
Vampire bat	vampire bat	19	103	10.8	0.64	[26]
A Nicaraguan community	Nicaraguan	108	1517	14.05	0.11	[27]
UC Irvine messages	UC Irvine	1302	19044	29.3	0.68	[28]
Online dating	POK	3562	18098	10.2	0.78	[41]

communities K with fivefold cross-validation, as in real data this value is usually unknown. Specifically, we perform edge prediction tasks using different values of K and evaluate the performance of CRAD on each data set by calculating the area under the curve (AUC) on a test set, after training the model on a training set. The value of K that yields the highest AUC is selected as the optimal number of communities.

Injecting anomalous edges. To evaluate the accuracy and precision of the model in detecting anomalous edges, we first need to know the true label of edges, being anomalous or regular. However, one of the challenges in this regard is the lack of data containing explicit anomalies. To address this challenge, we conduct an experiment where we inject n random edges between nodes in a real data set and label them as anomalous. We vary n to evaluate the impact of anomaly density $\rho_a = n/E$ on model performance. We then run our model on this manipulated data set and infer the expected value $\mathbb{E}[\sigma_{(ij)}] = \hat{Q}_{(ij)} \in [0, 1]$ for the edge labels, which also indicates the likelihood that the edges between two nodes are anomalous. Based on this, we assign labels to the edges. In this specific experiment, we label the first n pairs (i, j) with the highest values of $\hat{Q}_{(ij)}$ as anomalous edges. We measure the precision as a performance metric; this is the fraction of inferred anomalous edges which are correctly classified (in our case—since we fix the number of inferred anomalous edges to be equal to the number of injected anomalous edges—this also corresponds to recall, i.e., the fraction of true anomalous edges that are inferred as such).

To establish a benchmark comparison for evaluating the efficacy of our proposed algorithm, we conducted a comparative analysis of its performance against two naive classifiers, namely, uniformly random guess and majority class prior. The first makes a random prediction with a probability proportional to the percentage of anomalous edges; the latter always predicts the most common class in the data set. We computed various validation metrics, including AUC, accuracy, F1 score, and brief score, on results obtained using our model and these two naive classifiers. We find that the performance of these naive classifiers is suboptimal when applied to all data sets, and our proposed algorithm significantly outperforms both of them. A more detailed description of the results is presented in Appendix F.

1. Smaller data sets

a. Vampire bat network. The vampire bat network is a complex and dynamic social structure in which individual vampire bats form connections and share food with one another [26]. The bats have a remarkable ability to detect the body heat of other bats, even in complete darkness, allowing them to locate potential food sources and potential recipients

for food sharing. When a bat finds food, it will often regurgitate some of it and share it with other members of its network. This behavior, known as reciprocal altruism, is essential for the survival of the group, as it ensures that all members have access to food even when they are unable to find it themselves. The decision of who to feed is likely to be influenced by both the genetic relatedness of the individuals involved and their history of reciprocal sharing. Given this, we expect that reciprocity will play a significant role in determining which individuals form close social ties within this network. As such, when examining this data set, it will be important to carefully consider this effect. In our analysis, we use the data obtained from Ref. [26] and remove isolated nodes. The network consists of $N = 19$ nodes, $E = 103$ edges, and has high reciprocity of 0.64. In addition, we fix $K = 2$ as in Ref. [21].

As shown in Fig. 2, our model's ability to detect anomalies improves when there is a higher concentration of anomalies in the data set. The plot depicts the precision in detecting the anomalous edges, for a range of anomaly density, ρ_a . In a more specific case, Fig. 3 provides an example of how CRAD can be used for anomaly detection in the vampire bat data set. In this example, a set of edges with $\rho_a = 0.09$ was embedded in the system. In this figure, the entries of the estimated \hat{Q} matrix, which represent the probability of edges being anomalous, are categorized based on their true labels and assigned

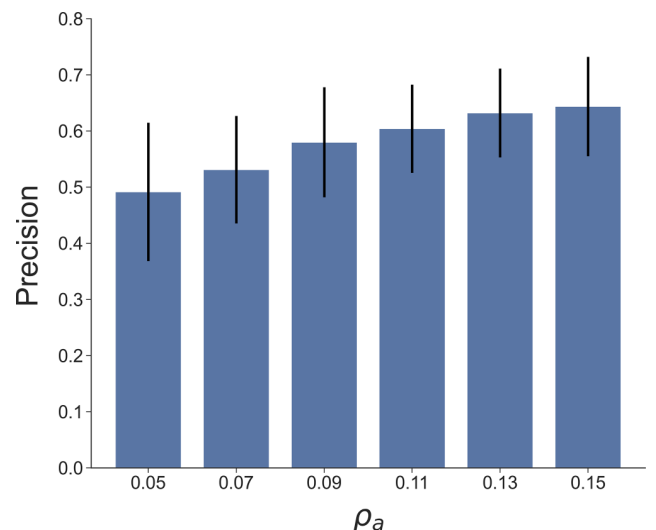


FIG. 2. Precision in detecting the injected edges in the vampire bat network. The precision increases by the increase in the number of anomalous edges injected in the network, i.e., anomaly density in the data set, ρ_a . The result is the average over ten randomly injected sets of edges; bars are standard deviations. Here we use the initialization $\pi = 0.1$.

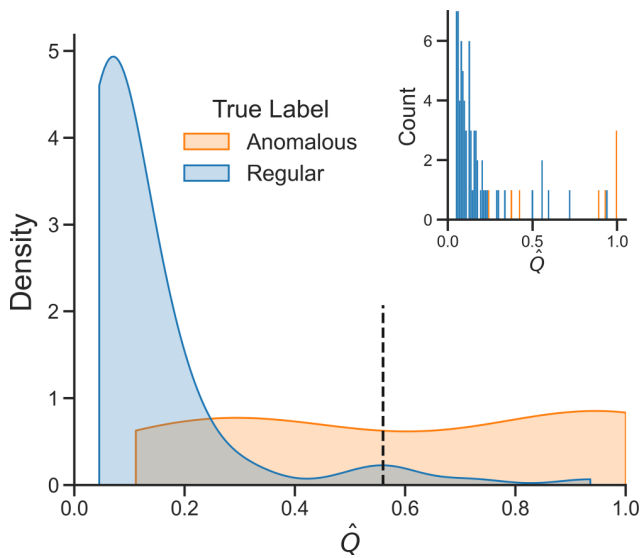


FIG. 3. Anomaly detection in the vampire bat network. We show the distribution of $\hat{Q}_{(ij)}$, i.e., the probability of a pair of edges (i, j) being anomalous, as estimated by CRAD. We distinguish the true regular and anomalous edges with different colors, blue and orange, respectively, to highlight their different inferred distributions. Here, $\rho_a = 0.09$ and $\pi = 0.1$. We measure a precision of 0.5. For this, we label as anomalous the fraction of ρ_a edges with highest $\hat{Q}_{(ij)}$. The vertical dashed line denotes the minimum $\hat{Q}_{(ij)}$ observed in this set of anomalous edges.

different colors to highlight their different inferred distributions of \hat{Q} . The plot clearly shows two different distributions, one with a high peak at $\hat{Q}_{(ij)} = 0$ and the other peaked around $\hat{Q}_{(ij)} = 1$. The inset reveals the presence of the second peak. Notably, the peak around $\hat{Q}_{(ij)} = 0$ extends up to 40 on the y axis in the inset, but the plot is truncated to highlight the more significant peak around $\hat{Q}_{(ij)} = 1$. The first corresponds to regular edges, which are thus correctly identified as such, while the latter are the injected anomalies, which are indeed assigned a higher probability of being anomalous. While there are few regular edges that have a high \hat{Q} , we observe that a significant density of anomalous edges is concentrated at $\hat{Q}_{(ij)} > 0.8$, indicating that the model is correctly assigning them as anomalous. Quantitatively, we measure precision and recall values of 0.5, obtained by labeling as anomalous the fraction of $\rho_a = 0.09$ edges with highest $\hat{Q}_{(ij)}$. Even though a small fraction of regular edges are classified as anomalous and vice versa, these numbers show that overall the algorithm is doing well at detecting the injected anomalies.

b. A Nicaraguan community. The next data set represents the social support network of indigenous Nicaraguan horticulturalists [27]. The original data set contains self-reported network data. Ties are reported by several individuals and these may be in disagreement with each other. Hence, we process it using the VIMuRe algorithm [16], which estimates probabilistically an underlying network structure from self-reported network data, provided by multiple reporters, accounting for reciprocity. The summary description of the estimated network by VIMuRe can be seen in Table I. In addition, it estimates the reliability $\theta > 0$ of each individual

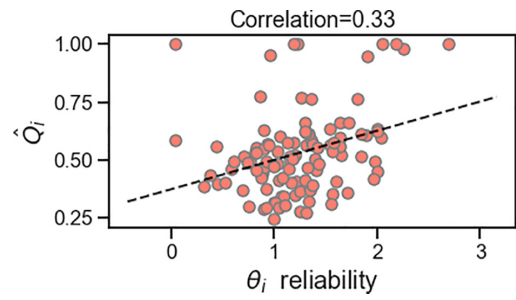


FIG. 4. Anomaly detection in a Nicaraguan social support network. We show a scatterplot of \hat{Q}_i (the maximum probability that one of the connecting ties of node i is anomalous), as estimated by CRAD, against θ_i , reporters' reliabilities, as estimated by the VIMuRe algorithm. The correlation is calculated as the Pearson coefficient; the dashed line is a linear fit to the data. Positive correlation signals that nodes that are more unreliable (high θ_i) tend to have an edge that is more likely to be labeled as anomalous among its connections.

reporter, with higher values denoting over-reporting. Reliabilities can be correlated to anomalies in that we expect that unreliable reporters may report nonexistence ties which we interpret as anomaly.

To assess this, we run VIMuRe twice. The first time, we run its default version and use it to collect estimates of reporters' reliabilities. The second time, we run it in a modified version where we fix the reliability parameters to a neutral value, assuming that all reporters are reliable. We use this output, the estimated network in this modified version, as input for CRAD. In this way, we aim at observing proxies for anomalous edges: These are some of the edges that involve unreliable reporters, as estimated in the first run of VIMuRe. Our model labels anomalies on edges; instead in this data set we have information on nodes (their reliabilities). We can build a correspondence between these two types of information by assuming that edges connected to the most unreliable reporters would have the highest value in the estimated \hat{Q} matrix. To quantify this match, we assign a value $\hat{Q}_i = \max_{j \in \partial i} \hat{Q}_{(ij)}$ to each reporter i , where ∂j is its neighborhood, being the maximum probability that one of its connecting ties is anomalous.

We expect \hat{Q}_i to be high for nodes that have a high unreliability θ . We find indeed a positive correlation of 0.33 between θ_i and \hat{Q}_i , as shown in Fig. 4. In particular, we observe that the edge (76,3) between the two most unreliable nodes has the maximum observed value of $\hat{Q}_{(ij)} = 1$, which is consistent with the findings reported in Ref. [16]. Notice that we expect this correlation to further increase if we were able to account explicitly for anomalies on nodes (instead of on edges). In this case, one could envision adapting our formalism to assign random variables σ_i to nodes, which may result in less tractable distributions and thus higher complexity, but may be more appropriate for applications in which nodes act consistently as anomalous. We leave this as an open question for future work.

2. Larger data sets

In this section, we test our algorithm on University of California (UC) Irvine and POK messages (from POK com-

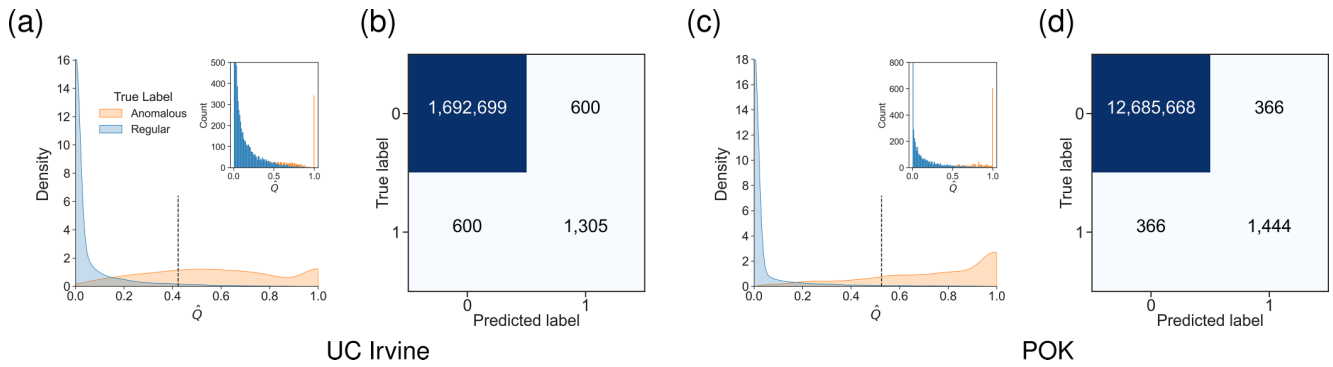


FIG. 5. Anomaly detection in the UC Irvine and POK networks. We show the distribution of $\hat{Q}_{(ij)}$, i.e., the probability of a pair of edges (i, j) being anomalous [(a), (c)] and the confusion matrix [(b), (d)] as estimated by CRAD, for the UC Irvine (left) and POK (right) data sets. We distinguish the true regular and anomalous edges with different colors, blue and orange, respectively, to highlight their different inferred distributions. Here, $\rho_a = 0.1$ and $\pi = 0.3$. We measure a precision of 0.68 for the UC Irvine and of 0.78 for the POK network. The vertical dashed line denotes the minimum $\hat{Q}_{(ij)}$ observed in this set of anomalous edges.

munity), as examples of larger data sets. In each case, we randomly select and add 10% additional edges, labeled as anomalous. The CRAD algorithm consistently produces reliable results in detecting anomalies in both data sets.

a. UC Irvine messages. The network of UC Irvine messages is composed of messages sent between users of an online community of students from the University of California, Irvine [28]. Each node in this communication network represents a user and each directed edge represents a message that was sent from one user to another. Our model consistently identified anomalies in this data set with a high level of accuracy, as evidenced by a particularly high peak in the distribution of $\hat{Q}_{(ij)}$ corresponding to anomalous edges in Fig. 5(a). The inset plot of Fig. 5(a) provides a more detailed view of the distribution of values around $\hat{Q}(ij) = 1$. Specifically, the plot reveals a clear peak in this region. Notably, the peak around $\hat{Q}(ij) = 0$ extends up to 12 500 on the y axis, but the plot is truncated to highlight the more significant peak around $\hat{Q}(ij) = 1$. This result is also quantified with a precision value of 0.68 in the confusion matrix shown in Fig. 5(b).

b. Network of online dating. The POK data set is a large data set containing the messages exchanged by users within the online dating POK community. The results depicted in Figs. 5(c) and 5(d) demonstrate the strong performance in identifying and reconstructing anomalous edges. Figure 5(c) illustrates how, also in this case, the distribution of \hat{Q} values for the anomalous edges is peaked around $\hat{Q} = 1$. The distribution of values around $\hat{Q}(ij) = 1$ is further analyzed in the inset plot of Fig. 5(c). In particular, the plot reveals a distinct peak in the vicinity of $\hat{Q}(ij) = 1$. Also in this case, we bounded the values shown in the y axis to emphasize the more prominent peak around $\hat{Q}(ij) = 1$, but values at $\hat{Q}(ij) = 0$ would otherwise extend up to 15 000. The precision value for the POK network is 0.78, as indicated by the corresponding entry in the confusion matrix in Fig. 5(d).

We can observe main common patterns in both UC Irvine and POK data sets: The distribution corresponding to regular edges is sharply peaked around $\hat{Q} = 0$ while the one for anomalous edges has a high peak around $\hat{Q} = 1$. This distinctive shape of the distribution of $\hat{Q}(ij)$ demonstrates the model's strong ability to distinguish regular edges from

anomalous ones. Taken together, these results support the efficacy of our classification methodology.

V. CONCLUSION

We introduce an expressive generative model to detect edge anomalies in networks that takes into account community membership and reciprocity as main mechanisms driving tie formation. By leveraging these two effects, it is able to detect what edges deviate from a regular behavior and estimate their probability of being anomalous. This inference is performed in a joint learning of edge anomalies and mixed memberships of nodes in communities, thus allowing practitioners to flag potential irregular edges while providing an interpretable community structure.

In contrast to common models for anomaly detection that rely on metadata on edges or nodes, our model takes as input only the adjacency matrix and estimates anomaly labels on the edges. It is an unsupervised model, meaning it does not require any input label to train it. These features make it particularly relevant in cases where extra information is not available—which is the case for many networked data sets—where the applicability of many machine learning methods for anomaly detection is significantly limited. As an example, traditional models for anomaly detection in financial transactions often rely on metadata such as transaction amount, location, and merchant information [5,29,30]. Instead, our model only requires the adjacency matrix of the transactions, which represents the connections between different account holders.

One key feature of our model is that it provides a joint probability for the existing pairs of edges between any pairs of nodes, allowing for the inclusion of reciprocity in the model, a relevant property in many directed networks. Furthermore, our model allows for mixed community membership, meaning that nodes can belong to more than one community. This is a more realistic representation of data structures compared to models that assume a single community membership for each node.

There are numbers of ways that our model could be further improved. As mentioned above, our model takes little

information in input, only the network's adjacency matrix and, optionally, the number of communities (which can otherwise be estimated from the input network with model selection criteria). A natural next step would be to extend the current model to account for extra information as node attributes, using ideas from generative models with both communities and attributes [31,32], or to consider techniques from semisupervised learning [33], in the case of availability of labels on a subset of the edges.

Furthermore, we can envision that, for rich and large data sets, deep learning architectures for anomaly detection [34–36] may be competitive methods. However, one could imagine extending standard architectures by combining them with the main ingredients of our model, in data sets where communities and reciprocity matter. The robust performance in detecting anomalies in real data with no extra information suggests that combining these insights with complex deep architectures may make the latter more expressive and thus boost predictive power.

Another type of extra information that is present in many real data sets is time [37]. Edges can be time-stamped and this could be used to improve estimates of anomalies. Hence, future work could be directed at generalizing our model to dynamical networks, for instance by combining insights from generative models for dynamic networks with communities [38–40].

It is important to note that the inferred labels for edges in our model should be treated as estimates rather than definitive conclusions. These labels should be used with caution in the study of a network, as further investigation may be necessary to fully understand the nature of anomalous edges. However, our model can provide valuable insights for practitioners to better understand and interpret the networks they are studying, especially when combined with their specialized knowledge and understanding of the data at hand.

ACKNOWLEDGMENTS

All the authors were supported by the Cyber Valley Research Fund. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting M.C.

APPENDIX A: DETAILED DERIVATIONS

Anomalous edges. As in the formation of anomalous edges, the reciprocated edges are independent. We apply the

condition $J_{(ij)}^{(a)} = 0$; therefore, from Eq. (4), we find

$$\frac{p_{11}^{(a)} p_{00}^{(a)}}{p_{10}^{(a)} p_{01}^{(a)}} = 1 \Rightarrow p_{11}^{(a)} = \frac{p_{10}^{(a)} p_{01}^{(a)}}{p_{00}^{(a)}}. \quad (\text{A1})$$

Moreover, $f_{(ij)}^{(a)} = f_{(ji)}^{(a)} = f^{(a)} \Rightarrow p_{10}^{(a)} = p_{01}^{(a)} = p^{(a)}$ and

$$f^{(a)} = \log \pi = \log \frac{p^{(a)}}{p_{00}^{(a)}} \Rightarrow p^{(a)} = \pi p_{00}^{(a)}. \quad (\text{A2})$$

Using the normalization condition, $p_{00}^{(a)} + p_{10}^{(a)} + p_{01}^{(a)} + p_{11}^{(a)} = 1$, and the results of Equation (A1) to (A2), we find the explicit mapping between the latent variables and the instances of $P^{(a)}(A_{ij}, A_{ji}|\theta_a)$ in Equation 2,

$$p_{00}^{(a)} = \frac{1}{Z_{ij}^{(a)}}, \quad p_{10}^{(a)} = p_{01}^{(a)} = \frac{\pi}{Z_{ij}^{(a)}}, \quad p_{11}^{(a)} = \frac{\pi^2}{Z_{ij}^{(a)}}, \quad (\text{A3})$$

where the normalization constant is

$$Z_{ij}^{(a)} = (1 + \pi)^2. \quad (\text{A4})$$

Regular edges. In order to find the explicit mapping between the latent variables and the instances of $P^{(r)}(A_{ij}, A_{ji}|\theta_r)$ in Equation 2, we follow the same procedure as in Ref. [21],

$$p_{01}^{(r)} = \frac{\lambda_{ji}}{Z_{(ij)}^{(r)}}, \quad (\text{A5})$$

$$p_{10}^{(r)} = \frac{\lambda_{ij}}{Z_{(ij)}^{(r)}}, \quad (\text{A6})$$

$$p_{11}^{(r)} = \frac{\eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}^{(r)}}, \quad (\text{A7})$$

$$p_{00}^{(r)} = \frac{1}{Z_{(ij)}^{(r)}}, \quad (\text{A8})$$

where the normalization constant is

$$Z_{(ij)}^{(r)} = \lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1. \quad (\text{A9})$$

Having these mappings, we can construct the marginal and conditional distributions of the ties. Thus, the marginal and conditional distributions of A_{ij} have the following densities, respectively:

$$P(A_{ij}) = \left[[p_{10}^{(r)}]^{A_{ij}} [p_{00}^{(r)}]^{(1-A_{ij})} + [p_{11}^{(r)}]^{A_{ij}} [p_{01}^{(r)}]^{(1-A_{ij})} \right] \times (1 - \mu) + \left[[p_{10}^{(a)}]^{A_{ij}} [p_{00}^{(a)}]^{(1-A_{ij})} + [p_{11}^{(a)}]^{A_{ij}} [p_{01}^{(a)}]^{(1-A_{ij})} \right] \times \mu, \quad (\text{A10})$$

$$P(A_{ij}|A_{ji}) = \frac{[p_{11}^{(r)}]^{A_{ij} A_{ji}} [p_{10}^{(r)}]^{A_{ij} (1-A_{ji})} [p_{01}^{(r)}]^{(1-A_{ij}) A_{ji}} [p_{00}^{(r)}]^{(1-A_{ij}) (1-A_{ji})}}{P(A_{ji})} \times (1 - \mu) + \frac{[p_{11}^{(a)}]^{A_{ij} A_{ji}} [p_{10}^{(a)}]^{A_{ij} (1-A_{ji})} [p_{01}^{(a)}]^{(1-A_{ij}) A_{ji}} [p_{00}^{(a)}]^{(1-A_{ij}) (1-A_{ji})}}{P(A_{ji})} \times \mu. \quad (\text{A11})$$

APPENDIX B: INFERENCE

Our goal is, given two mechanisms responsible for edge formation, first to determine the values of the parameters $\Theta = \{\{u_{ik}\}, \{v_{ik}\}, \{w_{kq}\}, \eta, \pi, \mu\}$, which determine the relationship between the anomaly indicator $\sigma_{(ij)}$ and the data, and then, given those values, to estimate the indicator $\sigma_{(ij)}$ itself.

We have the posterior:

$$P(\sigma, \Theta|A) = \frac{P(A|\sigma, \Theta)P(\sigma|\mu)P(\Theta)P(\mu)}{P(A)}. \quad (\text{B1})$$

Summing over all the possible indicators, we have

$$P(\Theta|A) = \sum_{\sigma} P(\sigma, \Theta|A), \quad (\text{B2})$$

which is the quantity that we need to maximize to extract the optimal Θ . It is more convenient to maximize its logarithm, log-posterior, as the two maxima coincide. We use Jensen's inequality:

$$L(\Theta) = \log P(\Theta|A) = \log \sum_{\sigma} P(\sigma, \Theta|A) \geq \sum_{\sigma} q(\sigma) \log \frac{P(\sigma, \Theta|A)}{q(\sigma)}, \quad (\text{B3})$$

where $q(\sigma)$ is a variational distribution that must sum to 1. In fact, the exact equality happens when

$$q(\sigma) = \frac{P(\sigma, \Theta|A)}{\sum_{\sigma} P(\sigma, \Theta|A)}. \quad (\text{B4})$$

This definition is also equivalent to maximizing the right-hand side of Eq. (B3) with respect to q .

Finally, we need to maximize the log-posterior with respect to Θ to get the latent variables. This can be done in an iterative way using the EM algorithm, alternating between maximizing with respect to q using Eq. (B4) and then maximizing Eq. (B23) with respect to Θ . In this work, we only fix priors for the σ_{ij} (Bernoulli distributions with parameter μ). For this variable we can thus estimate full posterior distributions; instead for the other parameters our model outputs point estimates. This could be modified by suitably specifying priors also for the reciprocity or community-related parameters. In this case, one could easily obtain maximum *a posteriori* (MAP) estimates with calculations similar to those reported here.

Defining $Q_{(ij)} = \sum_{\sigma_{(ij)}} q(\sigma_{(ij)}) \sigma_{(ij)}$, the expected value of $\sigma_{(ij)}$ over the variational distribution, we obtain

$$\begin{aligned} L(\Theta) = & - \sum_{\sigma} [q(\sigma) \log q(\sigma)] + \sum_{(i,j)} \{(1 - Q_{(ij)})(A_{ij} f_{ij}^{(r)} + A_{ji} f_{ji}^{(r)} + A_{ij} A_{ji} J_{(ij)}^{(r)} - \log Z_{(ij)}^{(r)}) \\ & + Q_{(ij)}((A_{ij} + A_{ji}) f^{(a)} - \log Z_{(ij)}^{(a)}) + Q_{(ij)} \log \mu + (1 - Q_{(ij)}) \log(1 - \mu)\}, \end{aligned} \quad (\text{B5})$$

and having Eqs. (7)–(10),

$$\begin{aligned} L(\Theta) = & - \sum_{\sigma} [q(\sigma) \log q(\sigma)] \\ & + \sum_{(i,j)} \left\{ (1 - Q_{(ij)}) \left(A_{ij} \log \sum_k u_{ik} v_{jq} w_{kq} + A_{ji} \log \sum_k u_{jk} v_{iq} w_{kq} + A_{ij} A_{ji} \log \eta \right. \right. \\ & \left. \left. - \log \left[\sum_{k,q} u_{ik} v_{jq} w_{kq} + \sum_{k,q} u_{jk} v_{iq} w_{kq} + \eta \sum_{k,q} u_{ik} v_{jq} w_{kq} \sum_{k,q} u_{jk} v_{iq} w_{kq} + 1 \right] \right) \right. \\ & \left. + Q_{(ij)}((A_{ij} + A_{ji}) \log \pi - 2 \log(\pi + 1)) + Q_{(ij)} \log \mu + (1 - Q_{(ij)}) \log(1 - \mu) \right\}. \end{aligned} \quad (\text{B7})$$

The derivative of the log-posterior with respect to η ,

$$\frac{\partial L(\Theta)}{\partial \eta} = \frac{1}{\eta} \sum_{(i,j)} (1 - Q_{(ij)}) A_{ij} A_{ji} - \sum_{(i,j)} (1 - Q_{(ij)}) \frac{\lambda_{ij} \lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \stackrel{!}{=} 0, \quad (\text{B8})$$

leads to a fixed-point equation,

$$\eta = f(\eta) = \frac{\sum_{(i,j)} (1 - Q_{(ij)}) A_{ij} A_{ji}}{\sum_{(i,j)} (1 - Q_{(ij)}) \left[\frac{\lambda_{ij} \lambda_{ji}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}, \quad (\text{B9})$$

which can be solved numerically with fixed-point methods. Alternatively, one can use root-finding methods to solve directly Eq. (B8) in η .

The equations for the remaining parameters need to be solved using Jensen’s inequality, and using $\log x < x$ to obtain $-\log x > -x$,

$$L(\Theta) \geq - \sum_{\sigma} [q(\sigma) \log q(\sigma)] \tag{B10}$$

$$+ \sum_{(i,j)} \left\{ (1 - Q_{(ij)}) \left(A_{ij} \sum_{k,q} \rho_{ijkq} \log \left(\frac{u_{ik} v_{jq} w_{kq}}{\rho_{ijkq}} \right) + A_{ji} \sum_{k,q} \rho_{jikq} \log \left(\frac{u_{jk} v_{iq} w_{kq}}{\rho_{jikq}} \right) \right. \right. \tag{B11}$$

$$\left. \left. + A_{ij} A_{ji} \log \eta - \left[\sum_{k,q} u_{ik} v_{jq} w_{kq} + \sum_{k,q} u_{jk} v_{iq} w_{kq} + \eta \sum_{k,q} u_{ik} v_{jq} w_{kq} \sum_{k,q} u_{jk} v_{iq} w_{kq} + 1 \right] \right) \right. \\ \left. + Q_{(ij)} ((A_{ij} + A_{ji}) \log \pi - 2 \log(\pi + 1)) + Q_{(ij)} \log \mu + (1 - Q_{(ij)}) \log(1 - \mu) \right\}, \tag{B12}$$

and the equality holds when

$$\rho_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k,q} u_{ik} v_{jq} w_{kq}}. \tag{B13}$$

We derive community parameters; for example, we start by considering u_{ik} ,

$$\frac{\partial L(\Theta)}{\partial u_{ik}} = \sum_j \left[(1 - Q_{(ij)}) \left[A_{ij} \sum_q \rho_{ijkq} \frac{1}{u_{ik}} - \sum_q v_{jq} w_{kq} - \sum_q \eta v_{jq} w_{kq} \lambda_{ji} \right] \right] \stackrel{!}{=} 0, \tag{B14}$$

and we finally obtain

$$u_{ik} = \frac{\sum_{jq} (1 - Q_{(ij)}) A_{ij} \rho_{ijkq}}{\sum_j \left[\frac{\sum_q (1 - Q_{(ij)}) (1 + \eta \lambda_{ji}) v_{jq} w_{kq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}. \tag{B15}$$

We find similar expressions for v_{ik} and w_{kq} :

$$v_{ik} = \frac{\sum_{jq} (1 - Q_{(ij)}) A_{ji} \rho_{jikq}}{\sum_j \left[\frac{\sum_q (1 - Q_{(ij)}) (1 + \eta \lambda_{ij}) u_{iq} w_{kq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}, \tag{B16}$$

$$w_{kq} = \frac{\sum_{i,j} (1 - Q_{(ij)}) A_{ij} \rho_{ijkq}}{\sum_{i,j} \left[\frac{(1 - Q_{(ij)}) (1 + \eta \lambda_{ij}) u_{ik} v_{jq}}{\lambda_{ij} + \lambda_{ji} + \eta \lambda_{ij} \lambda_{ji} + 1} \right]}. \tag{B17}$$

For π it yields the following:

$$\pi = \frac{\sum_{(i,j)} Q_{(ij)} (A_{ij} + A_{ji})}{\sum_{(i,j)} Q_{(ij)} (2 - A_{ij} - A_{ji})}. \tag{B18}$$

Similarly for μ ,

$$\frac{\partial L(\Theta)}{\partial \mu} = \sum_{(i,j)} \frac{1}{\mu} Q_{(ij)} - \frac{1}{1 - \mu} \sum_{(i,j)} (1 - Q_{(ij)}) \stackrel{!}{=} 0, \tag{B19}$$

yielding

$$\mu = \frac{1}{N(N - 1)/2} \sum_{(i,j)} Q_{(ij)}. \tag{B20}$$

To evaluate $q(\sigma)$, we substitute the estimated parameters inside Eq. (B4):

$$q(\sigma) = \frac{\prod_{(i,j)} \left[\frac{\exp\{(A_{ij} + A_{ji}) f^{(a)}\}}{Z_{(ij)}^{(a)}} \times \mu \right]^{\sigma_{(ij)}} \left[\frac{\exp\{A_{ij} f_{ij}^{(r)} + A_{ji} f_{ji}^{(r)} + A_{ij} A_{ji} J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} \times (1 - \mu) \right]^{1 - \sigma_{(ij)}}}{\sum_{\sigma_{(ij)}} \prod_{(i,j)} \left[\frac{\exp\{(A_{ij} + A_{ji}) f^{(a)}\}}{Z_{(ij)}^{(a)}} \times \mu \right]^{\sigma_{(ij)}} \left[\frac{\exp\{A_{ij} f_{ij}^{(r)} + A_{ji} f_{ji}^{(r)} + A_{ij} A_{ji} J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} \times (1 - \mu) \right]^{1 - \sigma_{(ij)}}$$

$$\begin{aligned}
&= \prod_{(i,j)} \frac{\left[\frac{\exp\{(A_{ij}+A_{ji})f^{(a)}\}}{Z_{(ij)}^{(a)}} \times \mu \right]^{\sigma_{(ij)}} \left[\frac{\exp\{A_{ij}f_{ij}^{(r)}+A_{ji}f_{ji}^{(r)}+A_{ij}A_{ji}J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} \times (1-\mu) \right]^{1-\sigma_{(ij)}}}{\sum_{\sigma_{(ij)}=0,1} \left[\frac{\exp\{(A_{ij}+A_{ji})f^{(a)}\}}{Z_{(ij)}^{(a)}} \times \mu \right]^{\sigma_{(ij)}} \left[\frac{\exp\{A_{ij}f_{ij}^{(r)}+A_{ji}f_{ji}^{(r)}+A_{ij}A_{ji}J_{(ij)}^{(r)}\}}{Z_{(ij)}^{(r)}} \times (1-\mu) \right]^{1-\sigma_{(ij)}}} \\
&= \prod_{(i,j)} Q_{(ij)}^{\sigma_{(ij)}} (1-Q_{(ij)})^{(1-\sigma_{(ij)})}, \tag{B21}
\end{aligned}$$

where

$$\begin{aligned}
Q_{(ij)} &= \frac{\exp[(A_{ij}+A_{ji})f^{(a)} - \log Z_{(ij)}^{(a)}] \mu}{\exp[(A_{ij}+A_{ji})f^{(a)} - \log Z_{(ij)}^{(a)}] \mu + \exp[f_{ij}^{(r)}A_{ij} + f_{ji}^{(r)}A_{ji} + J_{(ij)}^{(r)}A_{ij}A_{ji} - \log Z_{(ij)}^{(r)}] (1-\mu)} \\
&= \frac{\exp[(A_{ij}+A_{ji}) \log \pi - 2 \log(\pi+1)] \mu}{\exp[(A_{ij}+A_{ji}) \log \pi - 2 \log(\pi+1)] \mu + \exp[A_{ij} \log \lambda_{ij} + A_{ji} \log \lambda_{ji} + \log \eta A_{ij}A_{ji} - \log Z_{(ij)}^{(r)}] (1-\mu)} \\
&= \frac{\frac{\pi^{(A_{ij}+A_{ji})} \mu}{Z_{(ij)}^{(a)}}}{\frac{\pi^{(A_{ij}+A_{ji})} \mu}{Z_{(ij)}^{(a)}} + \frac{\lambda_{ij}^{A_{ij}} \lambda_{ji}^{A_{ji}} \eta^{A_{ij}A_{ji}} (1-\mu)}{Z_{(ij)}^{(r)}}}. \tag{B22}
\end{aligned}$$

Notice that this is exactly the expected value with respect to the variational distribution as previously defined.

Convergence criteria

The EM algorithm consists of randomly initializing u, v, w, η, π , and μ , then iterating Eqs. (B13), (B22), (B15)–(B17), (B9), (B18), and (B20), until the convergence of the following log-posterior:

$$\begin{aligned}
L(\Theta) &= \log P(\Theta|A) \geq \sum_{\sigma} q(\sigma) \log \frac{P(\sigma, \Theta|A)}{q(\sigma)} \\
&= -\sum_{\sigma} q(\sigma) \log q(\sigma) + \sum_{\sigma} q(\sigma) \{ \log P(A|\sigma; \Theta) + \log P(\sigma|\mu) \} \\
&= -\sum_{\sigma} q(\sigma) \log q(\sigma) + \sum_{\sigma_{(ij)}} q(\sigma_{(ij)}) \left\{ \sum_{(i,j)} [(1-\sigma_{(ij)})(A_{ij}f_{ij}^{(r)} + A_{ji}f_{ji}^{(r)} + A_{ij}A_{ji}J_{(ij)}^{(r)} - \log Z_{(ij)}^{(r)}) \right. \\
&\quad \left. + \sigma_{(ij)}((A_{ij}+A_{ji})f^{(a)} - \log Z_{(ij)}^{(a)}) + \sigma_{(ij)} \log \mu + (1-\sigma_{(ij)}) \log(1-\mu) \right\} \\
&= -\sum_{(i,j)} [Q_{(ij)} \log Q_{(ij)} + (1-Q_{(ij)}) \log(1-Q_{(ij)})] \\
&\quad + \sum_{(i,j)} \{ (1-Q_{(ij)})(f_{ij}^{(r)}A_{ij} + A_{ji}f_{ji}^{(r)} + A_{ij}A_{ji}J_{(ij)}^{(r)} - \log Z_{(ij)}^{(r)}) \\
&\quad + Q_{(ij)}((A_{ij}+A_{ji})f^{(a)} - \log Z_{(ij)}^{(a)}) + Q_{(ij)} \log \mu + (1-Q_{(ij)}) \log(1-\mu) \} + \text{const}, \tag{B23}
\end{aligned}$$

where we neglect const, constant terms due to the uniform priors. To calculate $q(\sigma)$, we used Eq. (B21), i.e., a Bernoulli distribution.

APPENDIX C: GENERATIVE MODEL

Being generative, the model can be used to generate synthetic networks with anomalies. For this, one should sample the latent parameters $\Theta = (u, v, w, \eta, \pi, \mu)$, then sample σ given the parameters. Finally, given the σ and the latent parameters, the adjacency matrix A could be constructed. For a given set of community parameters as the input parameters [15,24], the expected number of

anomalous and nonanomalous edges are $N^2 \mu \frac{\pi}{(1+\pi)}$, and $\mathbb{E}[M] = (1-\mu) \sum_{i,j} \frac{\lambda_{ij} + \eta \lambda_{ij} \lambda_{ji}}{Z_{(ij)}^{(r)}}$, respectively. Assuming a desired total number of edges E , we can thus multiply π, μ , and M by suitable sparsity constants that tune (i) the ratio of anomalous edges to the total number of edges, $\rho_a = N^2 \mu \frac{\pi}{(1+\pi)} / (N^2 \mu \frac{\pi}{(1+\pi)} + (1-\mu) \mathbb{E}[M]) \in [0, 1]$, and (ii) the success rate of anomalous edges π . Once these two are fixed, the remaining sparsity parameter for the matrix M

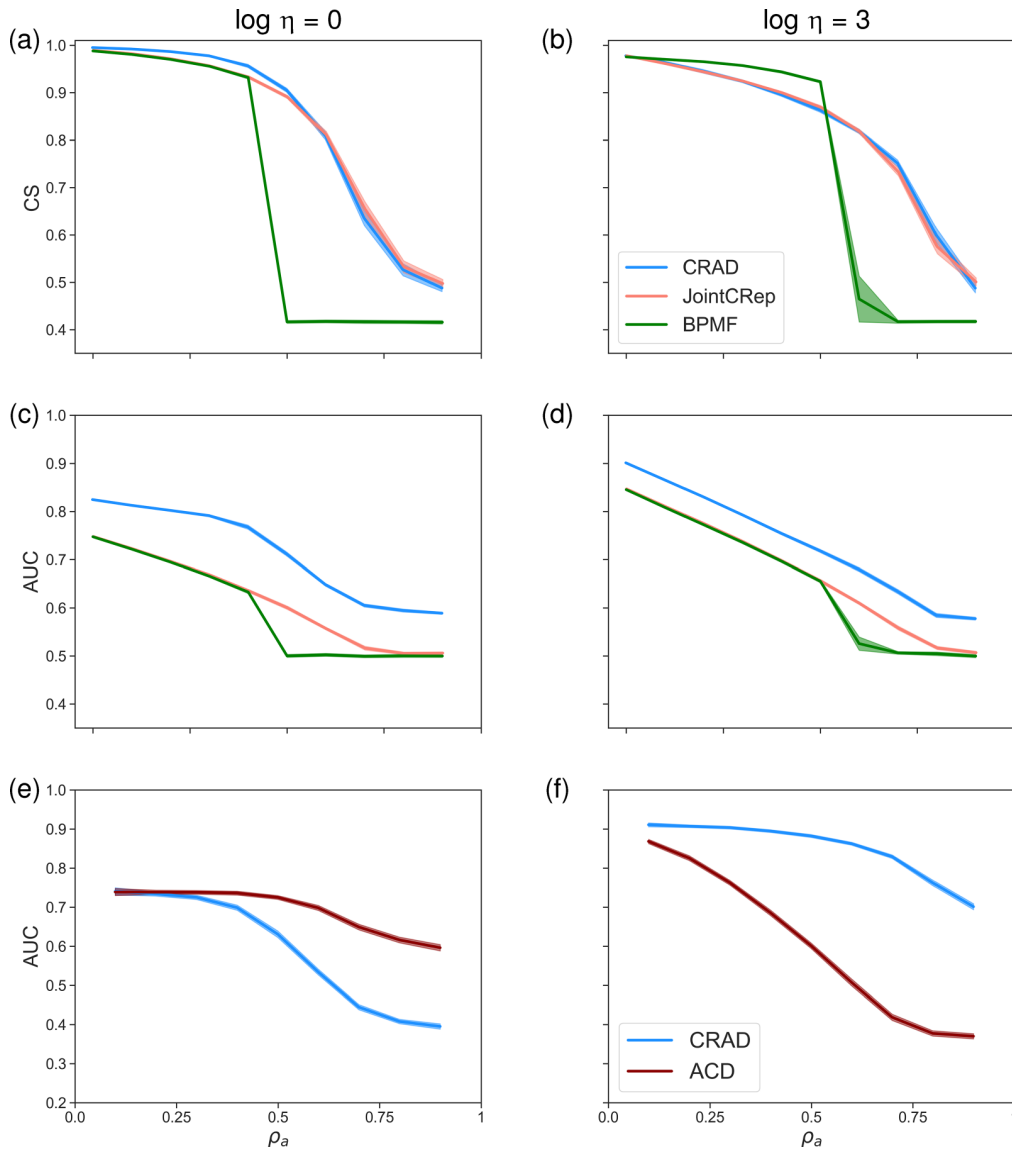


FIG. 6. Community detection, link prediction, and anomaly detection on synthetic network data sets. [(a), (b)] We compare the performance of CRAD against JointCRep and BPMF algorithms in community detection, as measured by cosine similarity (CS), and [(c), (d)] in link prediction tasks, as measured by AUC on held-out data. In addition, [(e), (f)] we test the ability to detect anomalies against a model that does not include a reciprocity effect (ACD), as measured by the AUC on a binary data set that contains what edges are regular and what are anomalous. The data sets have $N = 500$, average degree $\langle k \rangle = 60$, $K = 3$. The first column is for networks generated without reciprocity, $\log \eta = 0$, while the second column is for networks with positive reciprocity, $\log \eta = 3$. In the x axis we vary ρ_a , the ratio of anomalous edges over the total number of edges. Lines and shades around them are averages and standard deviations over ten network samples, respectively.

is estimated as

$$E(1 - \rho_a) = (1 - \mu) \sum_{i,j} \frac{\zeta \lambda_{ij} + \eta \zeta \lambda_{ij} \zeta \lambda_{ji}}{\zeta \lambda_{ij} + \zeta \lambda_{ji} + \eta \zeta \lambda_{ij} \zeta \lambda_{ji} + 1}, \tag{C1}$$

which can be solved with root-finding methods.

APPENDIX D: RESULTS ON SYNTHETIC NETWORKS

Figure 6 presents results of community detection, link prediction, and anomaly detection on data sets of synthetic networks.

APPENDIX E: REAL DATA: DATA-SET DESCRIPTION

Table I provides a summary of the key characteristics of the studied data sets. The data set of UC Irvine messages and online dating (POK) have undergone preprocessing that involved the removal of self-loops, retaining only nodes with both incoming and outgoing edges, and using only the giant connected components.

APPENDIX F: BENCHMARKING RESULTS AGAINST NAIVE CLASSIFIERS

To evaluate the classifiers, we used four commonly used performance metrics: AUC, accuracy, F1 score, and Brier

TABLE II. Performance metrics comparison. Results are averages and standard deviations in edge anomaly detection over ten different samples of sets of injected edges (10% of the total edges).

Data set	Metric	CRAD	Uniformly random guess	Majority class prior
vampire bat	AUC	0.789 ± 0.043	0.475 ± 0.086	0.500 ± 0.000
	F1 score	0.591 ± 0.084	0.052 ± 0.019	0.000 ± 0.000
	Accuracy	0.97507 ± 0.00511	0.49363 ± 0.02163	0.96952 ± 0.00000
	Brier score	0.02493 ± 0.00511	0.50637 ± 0.02162	0.03047 ± 0.00000
UC Irvine	AUC	0.873 ± 0.005	0.501 ± 0.005	0.500 ± 0.000
	F1 score	0.747 ± 0.007	0.002 ± 0.000	0.000 ± 0.000
	Accuracy	0.99943 ± 0.00000	0.50023 ± 0.00036	0.99888 ± 0.00000
	Brier score	0.00057 ± 0.00000	0.49977 ± 0.00036	0.00112 ± 0.00000
POK	AUC	0.895 ± 0.003	0.499 ± 0.006	0.500 ± 0.000
	F1 score	0.790 ± 0.006	0.000 ± 0.000	0.000 ± 0.000
	Accuracy	0.99994 ± 0.00000	0.50002 ± 0.00016	0.99986 ± 0.00000
	Brier score	0.000060 ± 0.000001	0.49998 ± 0.00000	0.00014 ± 0.00000

score. AUC is a measure of the classifier’s ability to distinguish between positive and negative samples. Accuracy measures the proportion of correctly classified samples. F1 score is the harmonic mean of precision and recall. For these three metrics, higher values indicate better performance. The Brier score measures the accuracy of probabilistic predictions, with lower values indicating better performance.

We compared the average performance of the proposed CRAD algorithm with two naive classifiers—“uniformly random guess” and “majority class prior”—on three real-world data sets: Vampire bat, UC Irvine, and POK. The “uniformly random guess” naive classifier makes random predictions with equal probabilities, but in the case of imbalanced data sets, the probability of each class is adjusted to match the per-

centage of the anomaly category. In this study, we imposed this bias to ensure that the number of anomalous and regular edges is correctly balanced. On the other hand, the “majority class prior” naive classifier always predicts the most common class in the data set. The averages are taken over ten samples with randomly injecting edges. In all these experiments, the number of injected edges is 10% of the total edges. The results in Table II show that the CRAD algorithm outperforms both naive classifiers across all data sets and in all performance metrics. Specifically, CRAD achieves higher AUC, accuracy, and F1 score values and lower Brier score values. These findings demonstrate the effectiveness of the proposed algorithm in accurately classifying the data, highlighting its potential for various real-world applications.

- [1] L. Akoglu, H. Tong, and D. Koutra, Graph based anomaly detection and description: A survey, *Data Min. Knowl. Discovery* **29**, 626 (2015).
- [2] V. J. Hodge and J. Austin, A survey of outlier detection methodologies, *Artif. Intell. Rev.* **22**, 85 (2004).
- [3] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, Network anomaly detection: Methods, systems and tools, *IEEE Commun. Surv. Tutorials* **16**, 303 (2014).
- [4] D. M. Hawkins, *Identification of Outliers (Monographs on Statistics and Applied Probability)*, Vol. 11 (Chapman & Hall, London, 1980)
- [5] E. Aleskerov, B. Freisleben, and B. Rao, Cardwatch: A neural network based database mining system for credit card fraud detection, in *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFER)* (IEEE, New York, 1997), pp. 220–226.
- [6] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey, *ACM Comput. Surv.* **41**, 1 (2009).
- [7] E. Eskin, Anomaly detection over noisy data using learned probability distributions, in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)* (Morgan Kaufmann, San Francisco, 2000), pp. 255–262.
- [8] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airolidi, A survey of statistical network models, *Foundations Trends Machine Learning* **2**, 129 (2010).
- [9] A. Bojchevski and S. Günnemann, Bayesian robust attributed graph clustering: Joint learning of partial anomalies and group structure, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 (AAAI Press, Palo Alto, California USA, 2018).
- [10] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, Anomaly detection in online social networks, *Social Networks* **39**, 62 (2014).
- [11] K. Henderson, T. Eliassi-Rad, C. Faloutsos, L. Akoglu, L. Li, K. Maruhashi, B. A. Prakash, and H. Tong, Metric forensics: A multi-level approach for mining volatile graphs, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)* (Association for Computing Machinery, New York, 2010), pp. 163–172.
- [12] V. Nikulin and T.-H. Huang, Unsupervised dimensionality reduction via gradient-based matrix factorization with two adaptive learning rates, in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, Proceedings of Machine Learning Research, Vol. 27, edited by I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver (PMLR, Bellevue, WA, 2012), pp. 181–194.
- [13] B. Perozzi and L. Akoglu, Discovering communities and anomalies in attributed graphs: Interactive visual exploration and summarization, *ACM Trans. Knowl. Discovery Data* **12**, 1 (2018).

- [14] S. Fortunato and D. Hric, Community detection in networks: A user guide, *Phys. Rep.* **659**, 1 (2016).
- [15] H. Safdari and C. De Bacco, Anomaly detection and community detection in networks, *J. Big Data* **9**, 122 (2022).
- [16] C. De Bacco, M. Contisciani, J. Cardoso-Silva, H. Safdari, G. Lima Borges, D. Baptista, T. Sweet, J.-G. Young, J. Koster, C. T. Ross, R. McElreath, D. Redhead, and E. A. Power, Latent network models to account for noisy, multiply reported social network data, *J. R. Stat. Soc. Ser. A* **186**, 355 (2023).
- [17] T. P. Peixoto, Reconstructing Networks with Unknown and Heterogeneous Errors, *Phys. Rev. X* **8**, 041011 (2018).
- [18] M. E. J. Newman, Estimating network structure from unreliable measurements, *Phys. Rev. E* **98**, 062321 (2018).
- [19] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Vol. 8 (Cambridge University Press, Cambridge, UK, 1994).
- [20] L. D. Molm, The structure of reciprocity, *Soc. Psychol. Q.* **73**, 119 (2010).
- [21] M. Contisciani, H. Safdari, and C. De Bacco, Community detection and reciprocity in networks by jointly modelling pairs of edges, *J. Complex Networks* **10**, cnac034 (2022).
- [22] H. Safdari, M. Contisciani, and C. De Bacco, Generative model for reciprocity and community detection in networks, *Phys. Rev. Res.* **3**, 023209 (2021).
- [23] B. Dai, S. Ding, and G. Wahba, Multivariate Bernoulli distribution, *Bernoulli* **19**, 1465 (2013).
- [24] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks, *Phys. Rev. E* **95**, 042317 (2017).
- [25] P. Gopalan, J. M. Hofman, and D. M. Blei, Scalable recommendation with hierarchical Poisson factorization in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, edited by M. Meila and T. Heskes (AUAI Press, 2015), pp. 326–335.
- [26] G. G. Carter and G. S. Wilkinson, Food sharing in vampire bats: Reciprocal help predicts donations more than relatedness or harassment, *Proc. R. Soc. B: Biol. Sci.* **280**, 20122573 (2013).
- [27] J. Koster, Family ties: The multilevel effects of households and kinship on the networks of individuals, *R. Soc. Open Sci.* **5**, 172159 (2018).
- [28] J. Kunegis, KONECT: The Koblenz Network Collection, in *Proceedings of the 22nd International Conference on World Wide Web* (Association for Computing Machinery, New York, 2013), pp. 1343–1350.
- [29] Y. Lucas, P.-E. Portier, L. Laporte, S. Calabretto, O. Caelen, L. He-Guelton, and M. Granitzer, Multiple perspectives HMM-based feature engineering for credit card fraud detection, in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (Association for Computing Machinery, New York, 2019), pp. 1359–1361.
- [30] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions, *Decis. Support Syst.* **75**, 38 (2015).
- [31] M. E. Newman and A. Clauset, Structure and inference in annotated networks, *Nat. Commun.* **7**, 11863 (2016).
- [32] M. Contisciani, E. A. Power, and C. De Bacco, Community detection with node attributes in multilayer networks, *Sci. Rep.* **10**, 15736 (2020).
- [33] X. Zhu and A. B. Goldberg, Introduction to Semi-supervised Learning, in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 3 (Springer, Berlin, 2009).
- [34] K. Liu, Y. Dou, Y. Zhao, X. Ding, X. Hu, R. Zhang, K. Ding, C. Chen, H. Peng, K. Shu, G. H. Chen, Z. Jia, and P. S. Yu, Pygod: A PYTHON library for graph outlier detection, [arXiv:2204.12095](https://arxiv.org/abs/2204.12095).
- [35] K. Liu, Y. Dou, Y. Zhao, X. Ding, X. Hu, R. Zhang, K. Ding, C. Chen, H. Peng, K. Shu, L. Sun, J. Li, G. H. Chen, Z. Jia, and P. S. Yu, Bond: Benchmarking unsupervised outlier node detection on static attributed graphs, *Advances in Neural Information Processing Systems* **35**, 27021 (2022).
- [36] E. Müller, P. I. Sánchez, Y. Mülle, and K. Böhm, Ranking outlier nodes in subspaces of attributed graphs, in *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)* (IEEE, New York, 2013), pp. 216–222.
- [37] N. Masuda and R. Lambiotte, *A Guide to Temporal Networks* (World Scientific, Singapore, 2016).
- [38] X. Zhang, C. Moore, and M. E. Newman, Random graph models for dynamic networks, *Eur. Phys. J. B* **90**, 1 (2017).
- [39] H. Safdari, M. Contisciani, and C. De Bacco, Reciprocity, community detection, and link prediction in dynamic networks, *J. Phys.: Complexity* **3**, 015010 (2022).
- [40] T. P. Peixoto and M. Rosvall, Modelling sequences and temporal networks with dynamic community structures, *Nat. Commun.* **8**, 582 (2017).
- [41] H. Makse, POK data set, <https://hmake.cuny.cuny.edu/>.

Latent network models to account for noisy, multiply reported social network data

Caterina De Bacco¹, Martina Contisciani¹, Jonathan Cardoso-Silva^{2,3},
Hadiseh Safdari¹, Gabriela Lima Borges⁴, Diego Baptista¹, Tracy Sweet⁵,
Jean-Gabriel Young⁶, Jeremy Koster^{7,8}, Cody T. Ross⁴, Richard McElreath⁴,
Daniel Redhead⁴ and Eleanor A. Power^{2,9}

¹Cyber Valley, Max Planck Institute for Intelligent Systems, Tuebingen, Germany

²Department of Methodology, London School of Economics and Political Science, London, UK

³Data Science Institute, London School of Economics and Political Science, London, UK

⁴Department of Human Behaviour, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

⁵Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA

⁶Department of Mathematics and Statistics and Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA

⁷Department of Anthropology, University of Cincinnati, Cincinnati, OH, USA

⁸Division of Behavioral and Cognitive Sciences, National Science Foundation, Alexandria, VA, USA

⁹Santa Fe Institute, Santa Fe, NM, USA

Address for correspondence: Eleanor A. Power, Department of Methodology, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: e.a.power@lse.ac.uk

Abstract

Social network data are often constructed by incorporating reports from multiple individuals. However, it is not obvious how to reconcile discordant responses from individuals. There may be particular risks with multiply reported data if people's responses reflect normative expectations—such as an expectation of balanced, reciprocal relationships. Here, we propose a probabilistic model that incorporates ties reported by multiple individuals to estimate the unobserved network structure. In addition to estimating a parameter for each reporter that is related to their tendency of over- or under-reporting relationships, the model explicitly incorporates a term for 'mutuality', the tendency to report ties in both directions involving the same alter. Our model's algorithmic implementation is based on variational inference, which makes it efficient and scalable to large systems. We apply our model to data from a Nicaraguan community collected with a roster-based design and 75 Indian villages collected with a name-generator design. We observe strong evidence of 'mutuality' in both datasets, and find that this value varies by relationship type. Consequently, our model estimates networks with reciprocity values that are substantially different than those resulting from standard deterministic aggregation approaches, demonstrating the need to consider such issues when gathering, constructing, and analysing survey-based network data.

Keywords: Social network data, mutuality, reliability, variational inference, latent network, network measurement

1 Introduction

Social network analysis has emerged as a fruitful framework for social scientists to represent and understand social relationships and their consequences (Borgatti et al., 2009). For example, patterns of interaction among people, as well as peoples' perceptions of their relationships, have been found to be important for their material wealth (Jackson, 2021), social position and

Received: December 21, 2021. Revised: July 27, 2022. Accepted: August 17, 2022

© (RSS) Royal Statistical Society 2023.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

welfare (Lin, 2002; Redhead & Power, 2022), and health and well-being (Holt-Lunstad et al., 2015; Perkins et al., 2015).

While new data sources now allow for the study of digitally mediated interactions (such as social media, mobile phone records, and other trace data; Eagle et al., 2009; Lazer et al., 2021; Park et al., 2018), social scientists' interest in day-to-day interactions and interpersonal relations are not always amenable to direct observation. Researchers, therefore, continue to rely on surveys where respondents identify the people with whom they have interactions or social relationships (Burt, 1984). A variety of approaches exist for eliciting self-reported network ties from respondents. Most common is the 'name generator' method, where respondents are asked to list the names of those with whom they have different types of relationships or interactions. Other approaches require a full roster, where respondents are asked about their relationship(s) with a set of possible partners (Marsden, 2005; Ross & Redhead, 2022; Warner et al., 1979).

Importantly, survey-based elicitations can be used not only for accounts of concrete interactions or exchanges, but can also facilitate a representation of respondents' subjective perceptions of their connections (Freeman, 1992; Krackhardt, 1987). Questions may be framed around more qualitative sentiments towards others—such as in friendships—and so do not merely document concrete interactions or observed events of exchange. For many substantive research questions, an individual's imperfect perception of their social relationships may be as (if not more) important as observable events of interaction or exchange. This has been highlighted by empirical research suggesting that individuals place considerable weight on their subjective relationships when making important decisions about who to cooperate with or support (Power, 2017; Redhead & von Rueden, 2021; von Rueden et al., 2019), and by work demonstrating that such relationships have strong associations with many important social and health-related outcomes (Kristiansen, 2004; Smith & Christakis, 2008).

The applicability of self-reported network data, however, has been subject to enduring debate within the social networks literature. Particularly when prompts query concrete exchanges or interactions, the quality of such data rests on the reliability of the self-reports that respondents provide, and numerous empirical studies have highlighted a plethora of potential biases in responses (Bernard et al., 1984; Killworth & Bernard, 1976). There is evidence that respondents' recall of their ties can be low, even over short periods of time (Brewer, 2000). For example, women within two West African communities were only able to accurately recall between 53% and 59% of their interactions across a 24-hr period prior to surveying (Adams et al., 2006). Alongside this, individual differences in the ability to recall ties may be predicted by relationship type, the number of partners a person has, and the duration of a given relationship (An, 2022; Bell et al., 2007). Both theoretical studies and empirically observed patterns of nominations suggest that individuals expressing particular attributes (e.g., high social status or power; Simpson et al., 2011) are more readily named, regardless of whether a relationship actually exists (Ball & Newman, 2013; Marin, 2004; Marineau et al., 2018; Redhead et al., Accepted; Shakya et al., 2017). The order in which questions appear within a survey, and the mode of elicitation, may further influence responses (Eagle & Proeschold-Bell, 2015; Pustejovsky and Spillane, 2009). That is, respondents have been shown to become fatigued, and report fewer relationships, when asked several name generator questions (Yousefi-Nooraie et al., 2019). Responses can also vary between interviewers, based in part on their attributes and their dynamic with the interviewee (Lungeanu et al., 2021; Marsden, 2003).

Noting all of these potential biases, one common practice is to obtain multiple reports on any single tie within a network. For relationships that are understood to be undirected, this is inherently captured with a single name generator question (i.e., both members of a friendship have the opportunity to report it). Previous research has found mixed results as to the concordance between respondents about the existence of their social relationships, with agreement in nominations ranging between 40% and 90% (Adams & Moody, 2007; Marsden, 1990). For relationships that are understood to be directed, multiple queries are necessary. One common approach is to 'double sample' a relationship, by asking respondents both who they go to for some type of assistance, and also who comes to them (Nolin, 2008). When combined with complete censusing of individuals, double sampling provides two perspectives on all relationships within a network, as both the giver and receiver have an opportunity to name their partner in each prompt. A recent survey of double-sampled network data has suggested that concordance between reporters is low, with an overall average of 10% agreement (Ready & Power, 2021).

Respondents need not be limited to reporting on the relationships in which they are directly involved, but may also be asked about the relationships between other individuals within the network. This type of data has been collected through ‘cognitive social structures’ roster designs—where respondents report on the relationships between all individuals within the network (Krackhardt, 1987; Newcomb, 1961)—though respondent fatigue means that this elicitation technique is somewhat uncommon. When it has been used, it has also shown relatively low levels of concordance between responses, highlighting that individual differences may guide respondents’ perceptions of their own relationships and the relationships of others (see Brands, 2013, for a review).

Low levels of inter-respondent concordance suggest that while having multiple reports on any relationship certainly provides new information, it does not necessarily resolve the issue of bias in reporting. Indeed, new issues may be introduced, if there are, for example, different reporting propensities for different queries. One key issue for double-sampled data, in particular, may be people’s expectation of, or desire for, mutually supportive, balanced relationships (Heider, 1958). The use of multiple prompts entailed in double sampling may lead to an inflation of apparent reciprocity, driven primarily by people’s propensity to name the same individuals across both prompts (Ready & Power, 2021). We use the term ‘mutuality’ here to refer to this apparent inflation of reciprocity. Overall, the low level of concordance found in multiply reported data raises the question of how to statistically account for the ambiguity introduced by conflicting reports of the same potential tie.

To examine the individual biases that shape self-reports of ties, and to estimate the effect of mutuality on the core properties of a network, we introduce a new latent network model for directed ties that is able to combine multiply reported network data, while accounting for the variable ‘reliability’ of respondents. Thus, we estimate a latent network, where the probability of an unobserved tie between two nodes is jointly dependent on the reports of multiple individuals and the reliability of those individuals. We validate our model by simulating noisy reports from a true network of ties, and then verify that we are able to recover the true generative network and the individual-level reporter reliability and mutuality parameters. Finally, we evaluate our model using two empirical datasets that feature double-sampled questions, one based on a ‘name generator’ design and the other based on a roster method design. We conclude by discussing our findings and outlining possible extensions of the model.

1.1 Related work

In the social sciences, simple deterministic rules are often used to aggregate multiple reports on what should nominally be the same relationship (Krackhardt, 1987; Lee & Butts, 2018). When data are collected via double sampling, for example, it is sometimes assumed that if one party forgets to report a relationship when asked (e.g., when they are asked who they give advice to), the other party may report that tie (e.g., when they are asked who they receive advice from). With such an expectation, the union of the two name generators is typically used (e.g., Nolin, 2010; Ready & Power, 2018). Alternatively, it could be assumed that relationships are only salient when they are mutually recognized; under such an expectation, the intersection of the two name generators would be preferred (e.g., Krackhardt & Kilduff, 1990). These aggregation rules rely on simple but strong expectations and presume consistency in how reporters respond to these questions. This, paired with the fact that the statistical tools used most frequently in the social sciences (e.g., exponential random graph models; Robins et al., 2007) assume that reported ties are a ‘true’ representation of a given network, can potentially lead to serious misrepresentations in the social relations of interest in a given study.

Several statistical methods have been proposed to resolve discordant reports for social network analysis (Butts, 2003; Holland et al., 1983; Kenny & La Voie, 1984; Killworth & Bernard, 1976; Redhead et al., Accepted; Sewell, 2019; Sosa & Rodríguez, 2021). Similar methods have also been introduced in other fields, like systems engineering (Amini et al., 2004), the biological sciences (D’haeseleer & Church, 2004; Hobson et al., 2021; Sprinzak et al., 2003), and physics (Newman, 2018b). Recently, for example, social scientists have attempted to tackle the problem of concordance by computing a ‘credibility score’ for every individual within a network, and determining whether a given tie exists based on each reporter’s assigned credibility (An & Schramski, 2015).

Considering this broad literature, we focus on methods most similar to our own, namely approaches that rely on an explicit generative model for reports that provide only imperfect

information about a true network of ties. For cognitive social structure data, in which each person reports on ties between every pair of people in the network, both Sewell (2019) and Sosa and Rodríguez (2021) have introduced models that aggregate network tie information across all reporters and simultaneously estimate error parameters for each reporter. The model proposed by Butts (2003) is more similar to our work in that it accommodates fewer reports on each tie and assumes the existence of a true underlying network. More recently, Redhead et al. (Accepted) introduce a latent network model for double-sampled data, which simultaneously estimates a true underlying network of directed ties and error parameters for each reporter, and directly incorporates mutuality. Our contribution involves an improved model for a latent network that accommodates *any number of reporters*, allows directed ties, and incorporates mutuality explicitly into the generative model of reports.

Our proposed model also requires a new estimation algorithm, which is an additional contribution of our work. This is because previous generative models for multiply reported data can be written as a finite mixture (Titterton et al., 1985) of probability distributions. For example, the probability distribution for a present tie could be different than the probability distribution for an absent tie. Finite mixture models can often be estimated with efficient algorithms, such as expectation-maximization, and have been used in network research where data come from unreliable reporters (Butts, 2003) or feature a significant amount of missingness (Peixoto, 2018). The unique formulation of our model requires an infinite mixture model approach and standard methods cannot be easily applied. Therefore, we propose a generative model for latent networks that simultaneously handles multiply reported ties and weighted reports, while allowing individuals to vary in reliability. To estimate our model, we introduce an efficient variational inference algorithm.

2 The model

Consider the problem of collecting a network of ties between individuals. These ties could, for instance, represent relationships commonly studied in the social sciences—such as loaning money, giving advice, or sharing food. This can be done by querying a set of M reporters about the existence of ties. The real network is not observed; responses of the reporters are the only observed data at our disposal. We assume that the unobserved network is correlated with these responses. Mathematically, we define this as an $N \times N$ -dimensional adjacency matrix, Y , where entries $Y_{ij} \in \{0, 1, \dots\}$ indicate the weight of the tie $i \rightarrow j$. For each tie type, the observed data is an $N \times N \times M$ -dimensional tensor, X , with entries X_{ijm} containing reports by respondent m about the tie $i \rightarrow j$.

We assume that each respondent can, in principle, report on any tie within the network. The exact rule of how reporters respond may change with the application, but may be flexibly represented by a binary mask, R , of entries R_{ijm} . We set $R_{ijm} = 1$ whenever a reporter, m , is surveyed about the possible existence of a tie from node i to node j , and set the entry to 0 otherwise. In scenarios where a network has been double-sampled—e.g., where the same reporter responds about *giving* and *receiving* social support—every tie type is sampled twice (for each reporter), once for each direction of the interaction. These binary masks are convenient in the inference procedure as they remove the contributions of nonreporters.

As an example, m can nominate who she gives advice to (giving) and who she receives advice from (receiving). In this case, $m \in \{i, j\}$, and we distinguish the direction of the reported data using the notation X_{ijm} to indicate i to j flows and X_{jim} to indicate j to i flows. While we gave an example for ties of type *advice*, the model applies for any type of directed tie. To keep the model flexible, we model weighted ties with positive and discrete weights, so that $X_{ijm} \in \{0, 1, \dots\}$. This also includes the binary case, when X_{ijm} captures only whether a tie exists or not.

One of the main objectives of our model is to estimate the structure of a latent network, Y , from the reported data, X . Note that the term ‘latent network model’ is also used for models predicting network ties that incorporate latent variables to account for tie dependence implicitly (e.g., latent space models; Hoff et al., 2002). In contrast, we are modelling networks whose ties are unobserved or latent. We adopt a probabilistic approach where we assume that X depends on Y in a potentially noisy way. This means that we infer a probability distribution over possible generative structures compatible with the reported ties. We assume *conditional independence* between the entries of X , given Y , and the model’s parameters. This is a common assumption made in network models (e.g., Newman, 2018b; Peixoto, 2018; Young et al., 2021), and makes estimation of the model more

tractable. Typical exceptions where this assumption may not hold are scenarios where an upper limit is set on the maximum number of nominations a reporter can make—e.g., when respondents are asked: *Who are your five closest friends?*. In these scenarios, there is a (weak) negative correlation between nominations, because the likelihood of future nominations is reduced each time a nomination is made by a respondent, simply because the respondent is strictly limited to an arbitrary, finite set of nominations (Hoff et al., 2013). While this is important to note, solving this problem is beyond the scope of the current manuscript.

A further core objective for our model is to estimate the reliability of reporters. Reporters may under-report (i.e., neglect to report a tie, when it does exist) or over-report (i.e., report a tie, when it does not exist), and we account for these biased reports by assigning a ‘reliability’ parameter, θ_m , to each reporter m . For ease of interpretability, we think of this parameter as a positive number taking higher values when the reporter exaggerates their reports and lower values when they under-report.

Finally, we incorporate the intuition that reporters tend to nominate the same people for both directions of a relationship, X_{ijm} and X_{jim} . We term this pattern ‘mutuality’, to keep the concept distinct from the standard concept of dyadic reciprocity (henceforth termed reciprocity) in the true unobserved network Y . Bringing all of these modelling consideration together, we posit that the expected value of the data can be given as

$$\mathbb{E}[X_{ijm} | Y_{ij} = k] = \theta_m \lambda_k + \eta X_{jim}, \tag{1}$$

where $\eta \geq 0$ is the mutuality parameter. Mutuality enters the model as an additive and positive contribution to the expected number of reported ties. This measures the possible increasing weight of a directed tie, given that we observe the same tie in the opposite direction, as reported by the same reporter. The parameter λ_k is a positive real value that needs to be inferred, which regulates the contribution of Y in determining X . Note that the index k here refers to the positive and discrete value posited for Y_{ij} . In case of binary entries, $k \in \{0, 1\}$, but in this work, we assume more generally $k \in \{0, 1, \dots\}$.

From this, we note how, for a given value of $\lambda_k > 0$, reporters with high θ_m tend to nominate more individuals, while reporters with smaller values tend to nominate fewer individuals. In contrast, a $\theta_m = 1$ indicates a neutral contribution (neither over-reporting nor under-reporting), hence we can interpret it as representing an unbiased reporter. Regardless of the reporter’s ‘reliability’, the existence of a tie X_{jim} in one direction increases the expected value of X_{ijm} in the opposite direction, when $\eta > 0$. This also implies that it may not be possible to identify the reliability of reporters that report a high percentage of ties in both directions, and in networks with high values of η . In these cases, in fact, the presence of a reported tie can be determined with a high likelihood based on the tie reported in the opposite direction.

To form a likelihood for the observed data that can accommodate various network and report structures—in particular, directed and weighted networks—we write the conditional distribution:

$$P(X_{ijm} | X_{jim}, Y_{ij} = k, \lambda_k, \theta_m, \eta) = \frac{(\theta_m \lambda_k + \eta X_{jim})^{X_{ijm}}}{X_{ijm}!} e^{-(\theta_m \lambda_k + \eta X_{jim})}. \tag{2}$$

Note that this choice of a Poisson distribution leads to an expected value for X_{ijm} as in equation (1). Furthermore, the positivity of the parameters makes this expression valid without the need of a link function. From this conditional, one can specify a two-point joint likelihood of (X_{ijm}, X_{jim}) by suitably defining the marginal distribution $P(X_{ijm} | Y_{ij} = k, \lambda_k, \theta_m, \eta)$. While there exist choices resulting in a consistent joint likelihood (see Section S1.3 for details), these may not result in simple, efficient closed-form updates of the parameters. Hence, we assume a pseudo-likelihood approximation (Besag, 1974) for the two-point likelihood, as done in Safdari et al. (2021)

$$\begin{aligned} &P(X_{ijm}, X_{jim} | Y_{ij} = k, Y_{ji} = q, \lambda_k, \lambda_q, \theta_m, \eta) \\ &\approx P(X_{ijm} | X_{jim}, Y_{ij} = k, \lambda_k, \theta_m, \eta) \times P(X_{jim} | X_{ijm}, Y_{ji} = q, \lambda_q, \theta_m, \eta). \end{aligned} \tag{3}$$

The model can be applied to any tie type encoded in the input data X , and it will output the reliability of a reporter for that tie type. One can potentially generalize this to a multi-layer framework by considering a unique θ_m for each reporter, regardless of tie type. This would then introduce a coupling between the reported X for various tie types, potentially increasing the complexity of the model. Alternatively, one could consider a different θ_m for each tie type. If these different types of reliability are considered independent from each other, then our model could be readily generalized to include these distinctions, without need for further extra coupling, but only additional distinct priors. This is essentially equivalent to running our model on each layer (i.e., tie type) individually, as we do in our numerical experiments on real data below.

Potentially, one could also include a different θ_m depending on the directionality of the ties—i.e., a θ_m^{\rightarrow} for ties sent and a θ_m^{\leftarrow} for ties received—capturing situations where reporters could over-report in one direction and under-report in another one. This would modify equation (3) to contain one of these two parameters inside the corresponding conditional distribution. If θ_m^{\rightarrow} and θ_m^{\leftarrow} are thought to be independent, so that their priors factorize, then this would lead to a straightforward generalization of the algorithm.

We assume that there are no contributions to the likelihood of X when a reporter is censored—i.e., when m is not given the chance to report on the tie $i \rightarrow j$. In empirical applications, this could arise, for example, when a survey design only asks about ties directly involving the reporter.

In addition to specifying the likelihood as in equation (2), we adopt a Bayesian approach and assume priors for the parameters and the unobserved Y . To maximize the flexibility of our model, we allow for positive and discrete values of Y by using a categorical prior

$$P(Y_{ij} = k; p_{ij}) = p_{ij,k}, \quad (4)$$

where p_{ij} is the parameter of the categorical prior distribution, and $\sum_k p_{ij,k} = 1$. The sum runs over the possible positive and discrete values of Y_{ij} . The resulting model can thus accommodate, for example, a binary network Y and weighted reports X , as the likelihood in equation (3) is valid for any number of values that Y can take. We then consider Gamma priors for the remaining parameters, as they are defined for positive real numbers, and are conjugate with the Poisson distribution, which makes calculations convenient.

3 Inference

Because of the possibility of mutuality in nominations, we do not have a closed-form joint distribution for (X_{ijm}, X_{jim}) , hence we consider the conditional distribution

$$\begin{aligned} P(\{X_{ijm}\}_m | \{X_{jim}\}_m, Y_{ij}, \lambda, \{\theta_m\}_m, \eta) &= \prod_m P(X_{ijm} | X_{jim}, Y_{ij}, \lambda, \theta_m, \eta) \\ &= \prod_k \left[P(Y_{ij} = k) \prod_m P(X_{ijm} | X_{jim}, Y_{ij} = k, \lambda_k, \theta_m, \eta) \right]^{Y_{ij,k}}. \end{aligned} \quad (5)$$

By using a pseudo-likelihood approximation as in [Safdari et al. \(2021\)](#), the full posterior can be written as

$$\begin{aligned} P(Y, \lambda, \theta, \eta | X) &\propto P(X | Y, \lambda, \theta, \eta) P(Y) P(\lambda) P(\theta) P(\eta) \\ &= \prod_{i,j} P(\{X_{ijm}\}_m | \{X_{jim}\}_m, Y_{ij}, \lambda, \{\theta_m\}_m, \eta) P(Y_{ij}; p_{ij}) \\ &\quad \prod_k P(\lambda_k; a_k, b_k) \prod_m P(\theta_m; \alpha_m, \beta_m) P(\eta; c, d) \\ &=: \mathcal{L}(\lambda, \theta, \eta, Y) \end{aligned} \quad (6)$$

$$=: \mathcal{L}(\lambda, \theta, \eta, Y) \quad (7)$$

where the proportionality results from the omission of an intractable normalization that does not depend on the parameters. To estimate the model, we use variational inference with a mean-field

variational family (Blei et al., 2017), which yields an approximate posterior distribution for the network and parameters. The algorithmic updates needed to find the best approximation to the posterior distribution follow a coordinate ascent routine, iteratively finding the best marginal posterior distribution of each parameter while holding the others fixed. We call the resulting algorithm VIMuRe, for Variational Inference for Multiply Reported data. The model is efficient, as it exploits the sparsity of the dataset. Specifically, the numerical implementation has a computational complexity that scales linearly with the number of nonzero entries of R , the reporters' mask, typically a sparse quantity. As a comparison, techniques based on sampling (e.g., HMC) can take an order of magnitude longer to run (see Blei et al., 2017), depending on the underlying complexity of the model. As the output results depend on the random initial configuration of the parameters, we run the algorithm several times and then consider the realization that resulted in the best ELBO value, as usually done in variational inference. This makes the output robust against initial values, as we expect a decreasing sensitivity to them for increasing $N_{\text{realisations}}$. In our experiments, we found that already $N_{\text{realisations}} = 5$ was a reasonable value to guarantee robust results. Pseudo-code for the algorithm is shown in Algorithm 1; see Section S1.1 for further details.

4 Simulation experiments

To validate our model, and study its performance in different regimes, we simulate synthetic data that reproduce our analysis scenarios—multiply reported network data that depend on a latent adjacency matrix—using the model itself. In detail, we first generate the network Y either with a flexible version of a mixed-membership stochastic block model (MULTITENSOR, De Bacco et al., 2017), a degree-corrected stochastic block model (DC-SBM, Karrer & Newman, 2011), or a probabilistic model with reciprocity (CRep, Safdari et al., 2021). We then generate the observed X given fixed reliability, mutuality, the generated network and the contribution of λ , collectively denoted by $\Theta = (Y, \theta, \lambda, \eta)$. We follow the approach described in Safdari et al. (2021), and for each reporter m we draw a pair (X_{ijm}, X_{jim}) consistently with the joint $P(X_{ijm}, X_{jim} | \Theta)$ in a two-step sampling routine, where we first generate one of the two reported ties and then the second one given the first, see Sections S1.2 and S1.3 for details.

In the simulations, we examine our ability to recover: (i) the underlying network, Y , and (ii) the individual reliabilities, θ_m . First, we generate synthetic networks reproducing three different scenarios. Two of these scenarios are extreme cases, where a fraction of reporters (θ_{ratio}) are tagged to be either over-reporters, or under-reporters, while all the others are reliable—i.e., they have $\theta_m = 1$, and their X entries are deterministically generated. In doing this, we document model performance in difficult cases, where the proportion of unreliable reporters is high. The third scenario, is more realistic. In this setting, we have both over- and under-reporters, as we draw θ_m from a Gamma distribution, providing a broad range of values. We vary the difference between λ_1 and λ_0 , such that the smaller this difference becomes, the noisier the problem gets, and thus the harder the inference tasks. Secondly, we investigate the ability of our model in recovering structural properties of the latent network Y —e.g., reciprocity, density, and communities—in other sets of synthetic networks.

In all experiments, we fit two versions of the model: a version with mutuality (VIMuRe_T) and a version without (VIMuRe_F). To provide a point of comparison, we also compute two baselines estimates of Y : (i) the *union*, in which a tie exists if at least one reporter reports that tie, and (ii) the *intersection*, in which all the reporters of a tie have to agree for the tie to exist. These two commonly used baselines represent the most and least inclusive approaches to integrating multiply reported data, and so provide reasonable comparisons for VIMuRe.

4.1 Results

We use the F1-score—i.e., the harmonic mean of precision (fraction of inferred ties that actually exist) and recall (fraction of existing ties found by the method) measures—to assess the ability of our model to recover Y , which is binary in our experiments. This choice is chiefly motivated by the fact that we have unbalanced data (since many fewer ties than possible tend to exist in empirical networks), that the F1 score is widely understood, and that our inferences based on F1 scores are qualitatively identical to those based on the Matthews correlation coefficient (Chicco & Jurman, 2020).

Algorithm 1: VIMuRe.

Input: Data X , Model \mathcal{L} , Variational family q .

Initialize the variational parameters γ , ϕ , ρ , ν to the priors with a small random offset.

while change in ELBO is above a threshold **do**

 For end each pair of nodes such that $X_{ijm} > 0$, update the multinomials:

$$\hat{z}_{mk}^1 \propto \exp\left\{\Psi(\gamma_m^{\text{shape}}) - \log \gamma_m^{\text{rate}} + \Psi(\phi_k^{\text{shape}}) - \log \phi_k^{\text{rate}}\right\}$$

$$\hat{z}_{ijm}^2 \propto \exp\left\{\Psi(\nu^{\text{shape}}) - \log \nu^{\text{rate}} + \log X_{ijm}\right\} = X_{ijm} \exp\left\{\Psi(\nu^{\text{shape}}) - \log \nu^{\text{rate}}\right\}$$

 where the proportionality is such that $\hat{z}_{mk}^1 + \hat{z}_{ijm}^2 = 1$.

 For each reporter, update the reliability parameters:

$$\gamma_m^{\text{shape}} = \alpha_m + \sum_{i,j,k} R_{ijm} \rho_{ij,k} X_{ijm} \hat{z}_{mk}^1$$

$$\gamma_m^{\text{rate}} = \beta_m + \sum_{i,j,k} R_{ijm} \rho_{ij,k} \frac{\phi_k^{\text{shape}}}{\phi_k^{\text{rate}}}.$$

 For each possible value k of Y_{ij} , update the parameters:

$$\phi_k^{\text{shape}} = a_k + \sum_{i,j,m} R_{ijm} \rho_{ij,k} X_{ijm} \hat{z}_{mk}^1$$

$$\phi_k^{\text{rate}} = b_k + \sum_{i,j,m} R_{ijm} \rho_{ij,k} \frac{\gamma_m^{\text{shape}}}{\gamma_m^{\text{rate}}}$$

 and:

$$\rho_{ij,k} \propto \exp\left\{\log p_{ij,k} + \sum_m R_{ijm} \left(X_{ijm} \hat{z}_{mk}^1 \mathbb{E}_{q(\lambda_k)}[\log \lambda_k]\right) - \frac{\phi_k^{\text{shape}}}{\phi_k^{\text{rate}}} \sum_m R_{ijm} \frac{\gamma_m^{\text{shape}}}{\gamma_m^{\text{rate}}}\right\}.$$

 Update the mutuality parameters:

$$\nu^{\text{shape}} = c + \sum_{i,j,k} \rho_{ij,k} \sum_m R_{ijm} X_{ijm} \hat{z}_{ijm}^2$$

$$\nu^{\text{rate}} = d + \sum_{i,j,m} R_{ijm} X_{ijm}.$$

end

Output: Variational parameters $(\gamma, \phi, \rho, \nu)$.

For readers more familiar with the latter, we include Matthews correlation coefficient results in [Supplementary Materials](#).

In the two extreme scenarios, where there are only over- or under-reporters, our model recovers the unobserved network, Y , better than approaches that take the union or intersection of the reported ties in X . The performance of our model is also more robust as the number of unreliable reporters and/or mutuality increases, see [Figure 1](#). In particular, our model with mutuality (VIMuRe_T) has a higher performance for high values of η , which is also a harder regime, as the performance of all methods decreases in this range. In general, the performance of the baselines decrease as the number of over- or under-reporters grows. For example, the union baseline estimates relationships that do not exist in the true network, when there are several over-reporters. Conversely, the intersection baseline underestimates the amount of ties, when a high fraction of individuals under-report. Our model overcomes these biases by accounting for reporters'

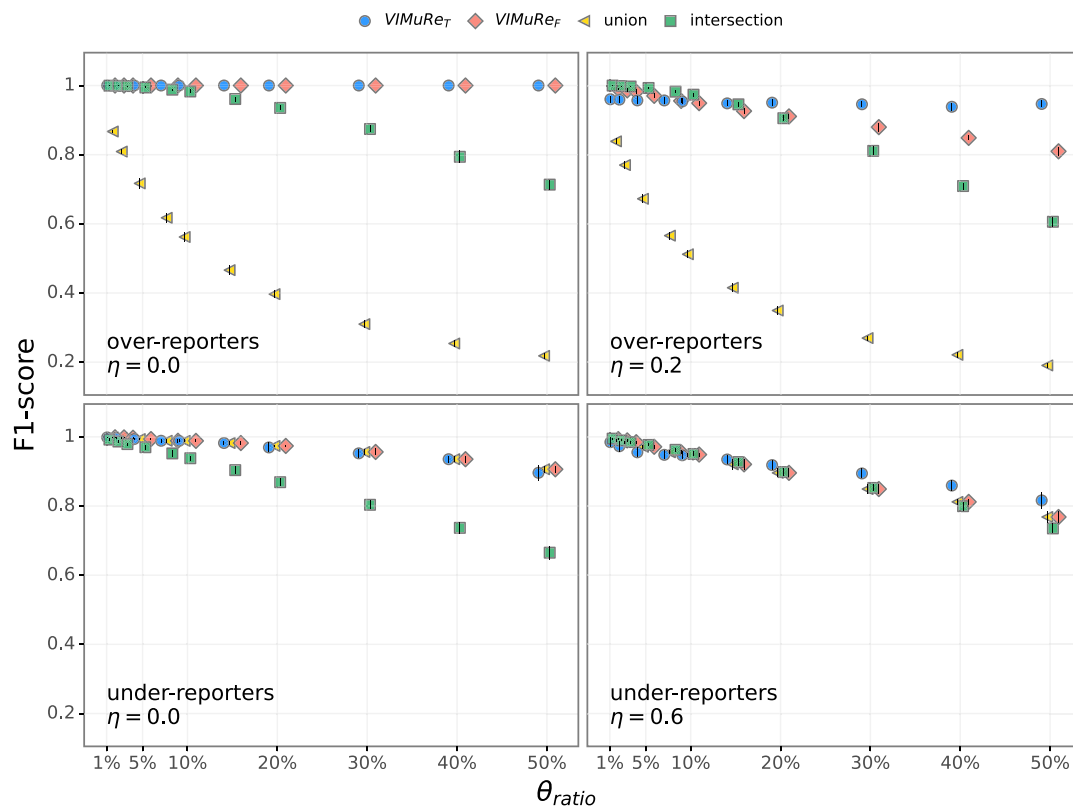


Figure 1. Estimating underlying network, Y , in synthetic networks with over- or under-reporters. Synthetic networks with $N = 100$ nodes and $M = 100$ reporters, generated with the benchmark generative model described in Sections S1.3 and 4, by varying the fraction θ_{ratio} of over-reporters (top) or under-reporters (bottom). The two columns represent networks generated without (left) and with (right) the mutuality effect η . The results are averages, and standard deviations calculated over ten independent synthetic networks. The accuracy of the estimate of the underlying network, Y , is measured with the F1-score. This measure ranges from 0 to 1, where 1 indicates perfect matching. See Figure S1 for similar plots based on the Matthews correlation coefficient and Figures S4 and S5 for additional experiments where M varies.

reliability, and this results in higher and more robust performance. However, when θ_{ratio} becomes too large, VIMuRe also fails since, as Figure S2 shows, recovering the reporters' reliability becomes harder. That said, the model with mutuality performs better at this task, and fails much more slowly than the model without mutuality, especially when η is large. To assess robustness in recovering Y as the number of reporters varies, we run further experiments keeping the same settings as above for $N = 300$ and varying $M \in [25, 300]$. For this simulation, we fixed $\theta_{ratio} = 0.50$, thus capturing the most challenging case explored in the original simulations in Figure 1. We find that while performance decreases as the number of reporters decreases, as expected, VIMuRe_T captures the ground truth of Y better than baseline implementations across different M , as shown in Figure S4.

Performance differences are more nuanced when we consider the more realistic experiment, which features a broad range of reporter reliabilities. F1-scores are lower than in the previous experiments, in general, and recovering the ground truth is particularly challenging when the difference between the mean number of reports of a tie being present and not, $\lambda_1 - \lambda_0$, is lower, see Figure 2. Intuitively, as the difference $\lambda_1 - \lambda_0$ decreases, both the zero and nonzero inputs of Y tend to make the same contribution in determining X ; thus, it becomes more difficult to distinguish true ties on the basis of reports. These experiments also further confirm what we observed in the previous experiments, that the hardest regime features the highest mutuality. A higher η means that a reporter will tend to nominate the same set of people for both *giving* and *receiving* questions, which results in X having less informative information. In these experiments, both versions of our model and the union baseline perform similarly while the intersection baseline performs much

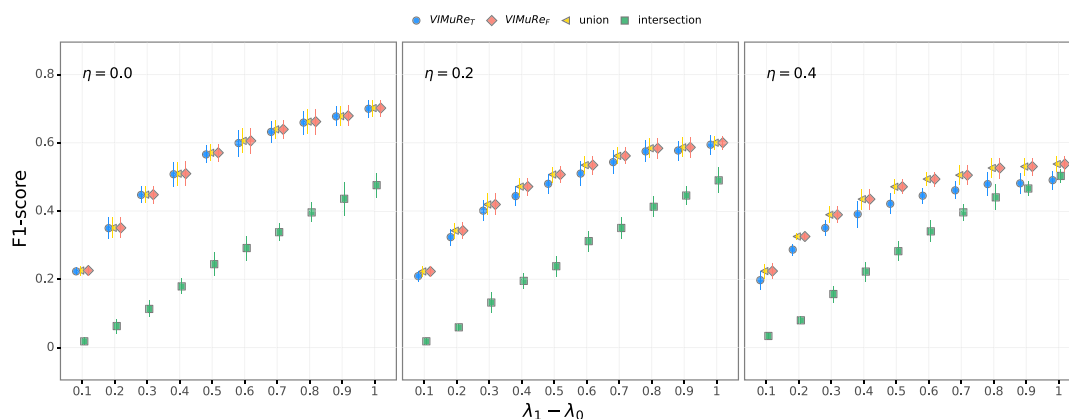


Figure 2. Estimating underlying network, Y , in synthetic networks with over- and under-reporters. Synthetic networks with $N = 100$ nodes and $M = 100$ reporters, generated with the benchmark generative model described in Sections S1.3 and 4, by varying the difference between λ_0 and λ_1 . The three columns represent networks generated with no (left), medium (centre), and high (right) mutuality, η . The results are averages and the standard deviations over ten independent synthetic networks. The accuracy of the estimate of the underlying network, Y , is measured with the F1-score. This measure ranges from 0 to 1, where 1 indicates perfect matching. See Figure S3 for similar plots based on the Matthews correlation coefficient.

more poorly. However, the performance gap decreases as mutuality increases and more ties are reported. Further studies on the recoverability of the latent network mutuality in synthetic experiments are provided in Section S1.4 and Figure S6.

Once we have an estimate, \hat{Y} , of the unobserved network, Y , a practitioner would be able to investigate structural properties in the latent network. To give an example, we will assess the ability of our model to capture reciprocity on the estimated \hat{Y} , a foundational feature of many social relations (Fehr & Gächter, 1998; Molm, 2010). To this end, we convert the posterior probability distribution $\rho_{ij,k}$ to a binary unweighted adjacency matrix. Since we considered binary data in our experiments, and thus $k \in \{0, 1\}$, we can obtain this by applying a threshold to the sub-tensor $\rho_{ij,k=1}$, as it represents the probability distribution of finding a $\hat{Y}_{ij} = 1$ entry. After each run of VIMuRe_T, we apply a range of thresholds $t_\rho \in [0.050, 0.075, \dots, 0.725, 0.750]$ such that we assign $\hat{Y}_{ij} = 1$ when $\rho_{ij,k=1} \geq t_\rho$, and keep track of the best t_ρ^* , for which reciprocity in the inferred network most closely matches the reciprocity of the ground truth. In Figure 3, we show that the relationship between this optimal threshold and the mutuality, η_{est} , as inferred by VIMuRe_T, can be approximated by the linear equation

$$t_\rho^* = 0.33\eta_{\text{est}} + 0.10. \quad (8)$$

In fact, in Figure 4, we show that VIMuRe_T outperforms all other models at this task when the threshold on ρ was set according to the heuristic proposed above. Reciprocity estimated by the model was a closer match to the reciprocity of the true unobserved network even in simulations with high values of mutuality, a scenario where other methods tend to overestimate reciprocity. Density of the inferred network is also closer to ground truth, when compared with baseline methods in most scenarios, as can be seen in Figure S7. These results mean that despite the small gap on the value of F1-score, VIMuRe_T may be able to provide a good estimate of structural properties of Y . The values of t_ρ^* in equation (8) are valid for the settings considered in the experiments analysed here and reported in Table S1, which control how synthetic networks are generated. Being a heuristic, one can in principle obtain a different formula when simulating data under different assumptions (e.g., varying N , M or reciprocity).

As a final test, we also show that VIMuRe allows the underlying community structure of a network to be recovered, even when it is measured noisily. To this end, we use a latent network that has planted overlapping communities, and generate reports as before. We then estimate the network, and finally recover communities using a probabilistic generative model with latent variables (De Bacco et al., 2017). Figure 5 shows the result of this experiment, and illustrates that VIMuRe is

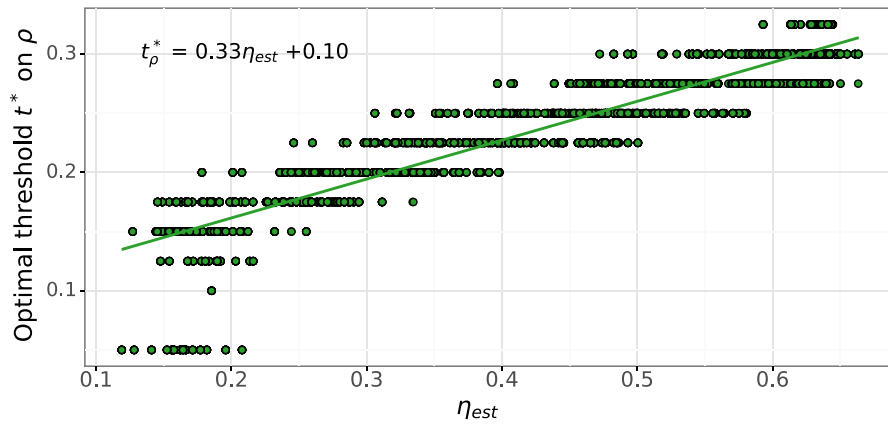


Figure 3. Synthetic networks with $N = 100$ nodes and $M = 100$ reporters, generated with the benchmark generative model described in Sections S1.3 and 4 with $\lambda_1 - \lambda_0 = 1.0$, and planted reciprocity values around ≈ 0.2 on the ground truth network, Y . The plot shows that the threshold that best captures reciprocity is linearly correlated with η_{est} .

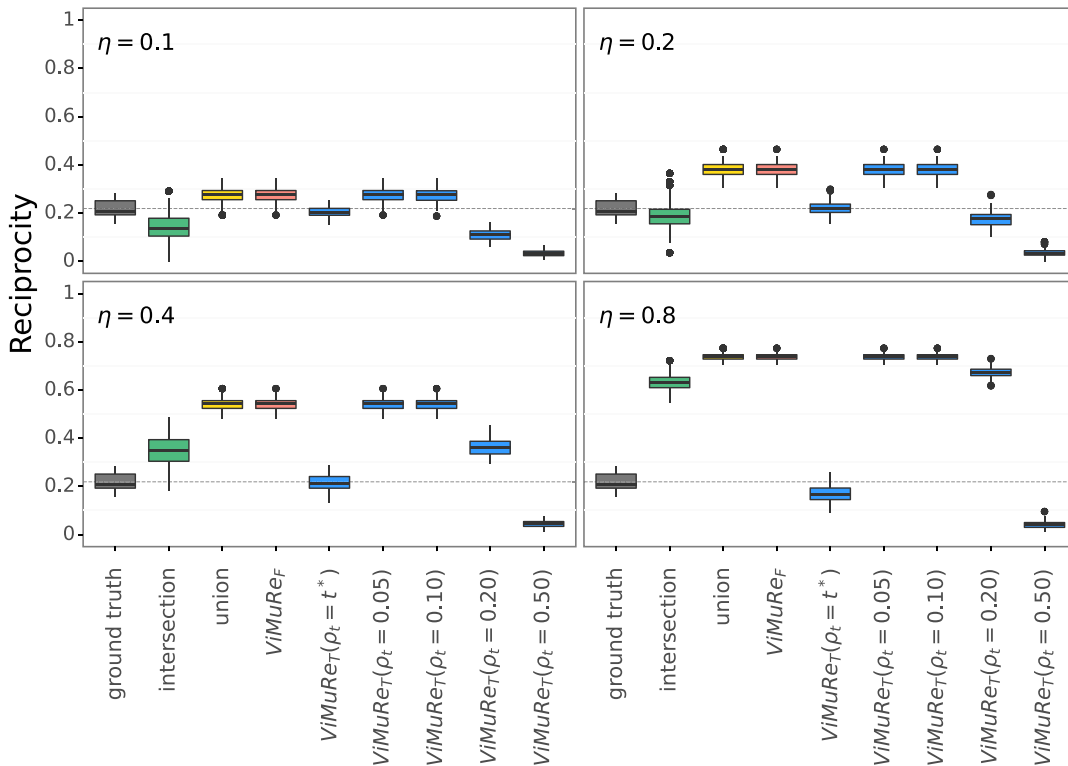


Figure 4. Reciprocity recovery from synthetic networks with $N = 100$ nodes and $M = 100$ reporters, generated with the benchmark generative model described in Sections S1.3 and 4 with $\lambda_1 - \lambda_0 = 1.0$, and planted reciprocity values around ≈ 0.2 (horizontal dashed line) on the ground truth network, Y . The four sub-plots represent networks generated with low (top left), medium (top right), to increasingly high (bottom left and right) mutuality effects, η . The box plots are distributions of the reciprocity in Y , over a sample of one hundred synthetic networks.

more robust than other approaches across different mutuality values, providing slightly better results than all other models; the intersection performs the worst. The qualitative example on the right panel of Figure 5 highlights how ViMuRe_T infers a partition closer to the ground truth than those inferred by the other methods, especially when mutuality is higher.

To summarize, our simulation experiments suggest that the use of a generative model with latent variables results in more robust estimates of the true underlying network, Y , in comparison to

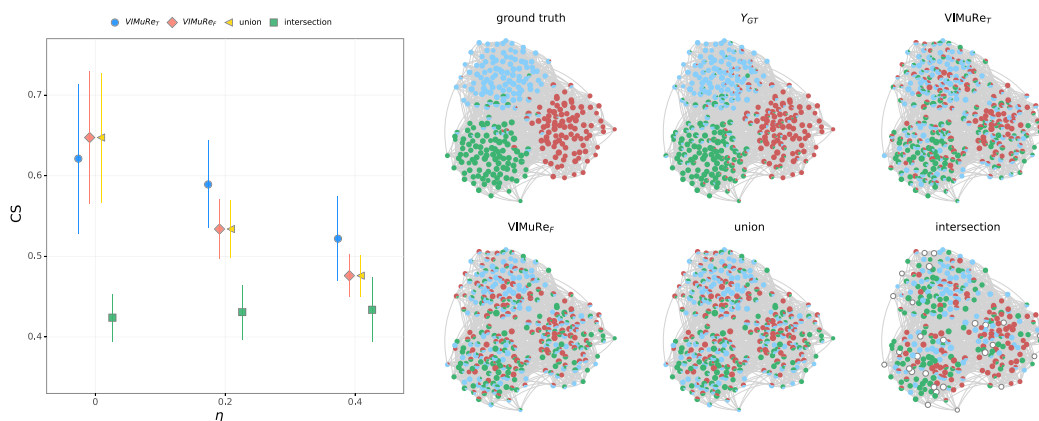


Figure 5. Community structure recovery from synthetic networks with both over- and under-reporters. Synthetic networks with $N = 300$ nodes and $M = 300$ reporters, generated with the benchmark generative model described in Sections S1.3 and 4 with $\lambda_1 - \lambda_0 = 1.0$. The results shown in the left frame are averages and standard deviations over ten samples of synthetic networks, generated by varying the mutuality parameter, η . The accuracy in recovering the overlapping community structure is measured with the cosine similarity (CS) using the inferred membership vectors. In the right frame, we plot examples of the partitioning of a synthetic network, generated with $\eta = 0.2$. The ‘ground truth’ is the partition used to generate Y , and Y_{GT} stands for the partition found by using the true Y . Nodes colored in white represent isolated nodes.

deterministic approaches (such as taking the union or intersection of sub-tensors). Furthermore, our model yields an estimate of reporter reliabilities, which can provide additional insights about the data-generating process. In addition, we note that our model performs better than other models when we include the mutuality parameter, η . In particular, VIMuRe_T shows better results in estimating reciprocity than VIMuRe, specifically in cases where people’s propensity to report mutuality in their relationships is high.

5 Analysis of Nicaragua data

We apply our modelling approach to data collected from a horticulturalist community in Nicaragua (see Koster, 2018, for more detail on the population and measurement instruments). These data were collected using a roster-based design, where all adult residents within the community were presented with a list of all other adult residents, and were asked two questions about relationships related to social support (i.e., *Who provides tangible support to you at least once per month?* and *Who do you provide tangible support at least once per month?*). Previous studies have performed separate analyses on the two questions (Koster, 2018; Simpson, 2022). We examine both questions in a single model, examining the potential biases that shape the reports of social support.

In this dataset, the reports vary significantly across reporters, with some reporters nominating many ties and others nominating fewer. It is, therefore, reasonable to have the priors on θ_m reflect this. We can incorporate this insight by running the inference in two steps, where we first run VIMuRe with a weak prior that is the same for all reporters, while in a second step we run VIMuRe with a prior proportional to the posterior mean of θ_m inferred in the first step. This is in line with Empirical Bayes approaches (Casella, 1985; Morris, 1983; Robbins, 1955) that estimate prior distributions from the data. This approach allows us to obtain a wider range of reliabilities so that we can better distinguish possible exaggerators than when using the same prior for all reporters.

Applying VIMuRe produces estimates of a network—which is binary and obtained by applying the optimal threshold in equation (8)—that has properties (e.g., mean degree, reciprocity) that fall somewhere between the results of taking the union (which returns an incredibly dense network) and taking the intersection of the double-sampled ties (which returns an extremely sparse network). See Table S2, for a summary. Overall, mutuality was estimated to be $\eta_{\text{est}} = 0.540$ and reciprocity was 0.11.

In contrast to other survey data such as name generators, where survey design may contribute to under-reporting, the roster-based design makes it much easier for respondents to make many nominations. In the roster design in Nicaragua, informants reported approximately 25 alters for each prompt, substantially more than the average of 4 from the constrained name generators used in the

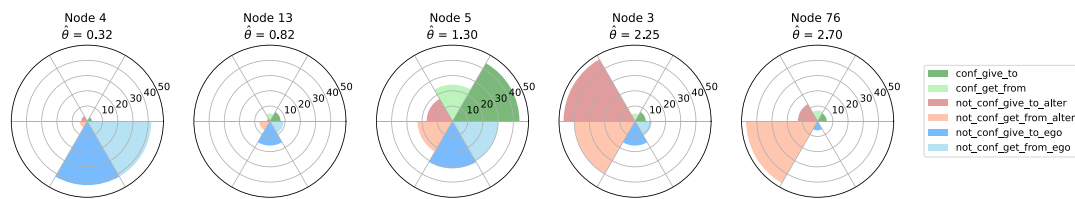


Figure 6. Example of individual reliabilities. Pie plots show six different configurations for the reported ties (two per each direction of a tie): ties confirmed by both reporters (conf ‘give to’, conf ‘get from’); ties reported by m but not confirmed by others (not conf ‘give to’ (alter), not conf ‘get from’ (alter)); ties reported by others but not by m (not conf ‘give to’ (ego), not conf ‘get from’ (ego)). Each plot is a different reporter; their estimated reliability $\hat{\theta}$ is printed on top. Each slice of the pie is one tie reported in one the six possible ways, represented by the colours. In this example, we consider reporters from the Nicaragua dataset.

study described below in Section 6 (see Table S2 for full network statistics). As can be seen in Figure 6, however, we nevertheless observe reporters with low θ_m , who were nominated by several others, but nominated relatively few themselves (e.g., Nodes 4 and 13). On the opposite extreme, we see reporters with high θ_m , who nominated many others, but whose ties are not confirmed by those alters (e.g., Nodes 3 and 76). In between, we show an example of a reporter (Node 5) with intermediate value of θ_m , who nominates several others in a way consistent with the reports of others. The distribution of reliability for reporters in this dataset can be seen in Figure S8.

Since these data are not explicitly generated with the generative model assumed by VIMuRe, we run a goodness-of-fit test to ensure that the model is appropriate for the above analysis. To do this, we use a series of posterior-predictive checks (Gelman et al., 1996, 2013), which compare the Nicaragua data with synthetic data \tilde{X} generated using the fitted model. The posterior-predictive distribution is defined as

$$P(\tilde{X} | X) = \sum_Y \iiint P(\tilde{X} | Y, \lambda, \theta, \eta) P(\lambda, \theta, \eta, Y | X) d\lambda d\theta d\eta \tag{9}$$

and one can generate samples from this distribution by first obtaining samples $(Y, \lambda, \theta, \eta)$ from the variational approximation to the posterior distribution, and then using these parameters as input to create new synthetic data, \tilde{X} , from the likelihood described in Section 2. A good fitted model should lead to new synthetic data \tilde{X} that resembles the input X . We run two numerical posterior-predictive tests to assess the appropriateness of the VIMuRe model: (i) a direct comparison between the elements of the Nicaragua data X and the distribution of \tilde{X} , and (ii) a test checking whether two samples from the posterior-predictive distribution are typically more, equally, or less distant from one another than a sample from the posterior-predictive distribution and the Nicaragua data (Young et al., 2021). The results shown in Figure 7 confirm that the VIMuRe model is appropriate for our analysis.

6 Analysis of social support networks in Karnataka

To further highlight the broad applicability of our modelling approach across elicitation methods, we apply VIMuRe to a dataset of social support networks collected from 75 villages in the Indian state of Karnataka (Banerjee et al., 2013). As part of a larger project, a series of name generators were asked of most of the adult members of a subset of households in each village (overall, about 46% of all households were surveyed). The name generators included questions about four double-sampled relationships: who people give advice to or receive advice from (*Advice*), who people would borrow from or lend a small amount of money to (*Money*), who people go to or receive as visitors (*Visit*), and who people would borrow kerosene and rice from or lend kerosene and rice to (*Household Items – ‘HH Items’ in the plots*). In the past, these data have been studied by aggregating responses from multiple household respondents and taking the union of the double-sampled questions (Banerjee et al., 2013; Jackson et al., 2012).

Our results suggest that the reciprocity values in these networks are in fact lower than what would be obtained by simpler approaches, as shown in Figure 8, generally, and illustrated for one specific village and tie type in Figure 9. While we do not have ground truth values in this

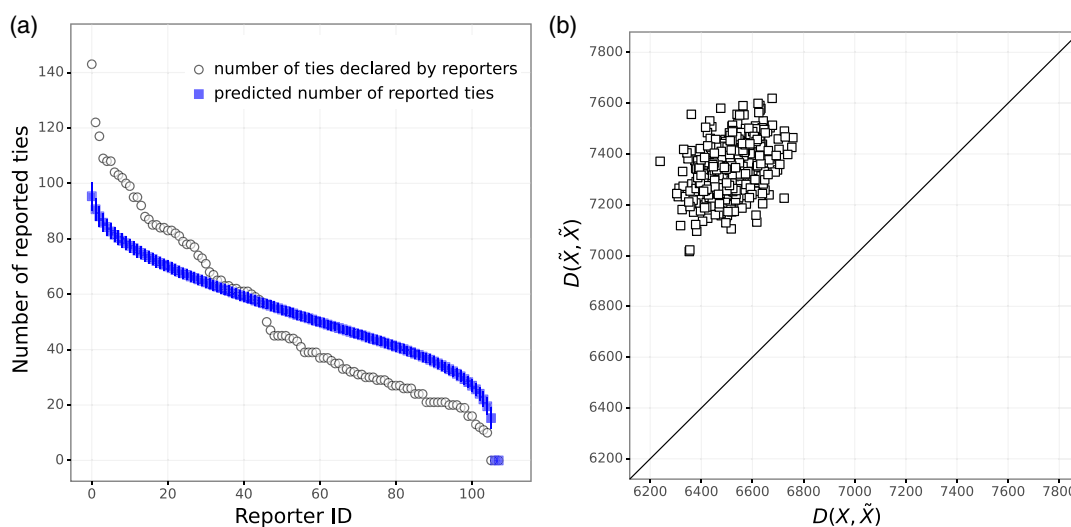


Figure 7. Example goodness-of-fit analysis for the Nicaragua dataset. (a) Number of ties declared by each reporter (squares) across the double-sampled social support question, compared with the average predicted number of reported ties (circles). Error bars correspond to standard deviations computed with $n = 500$ samples from the posterior distribution. (b) Scatter plot showing a model-model ($D(\hat{X}, \hat{X})$) versus model-data ($D(X, \hat{X})$) comparison. Each dot corresponds to one posterior-predictive sample and illustrates the distance between this sample and the Nicaragua data on the horizontal axis (model-data), and another random posterior-predictive sample on the vertical axis (model-model). The model can be deemed appropriate when these two distances are similar—i.e., if the scatter plot is a point-cloud centred on or close to the diagonal (Young et al., 2021). We selected the Hamming distance D as the test statistics, defined as the number of pairs of entries (X_{ijm}, \hat{X}_{ijm}) in disagreement between two datasets.

case, we note that these numbers are similar to those obtained on the synthetic networks in our experiments shown in Figure 4. In particular, they mimic the situation with high mutuality, where the union and intersection significantly overestimate the reciprocity on Y , whereas VIMuRe identifies the correct range of values. These results suggest that reciprocity will likely be overestimated in double-sampled network data, when the reports have high mutuality. Of the four tie types in the data from Karnataka, the estimations made by VIMuRe suggest that the ‘Advice’ layer has the lowest reciprocity values on average (0.39 ± 0.07), with ‘Money’ (0.43 ± 0.08), ‘Visit’ (0.47 ± 0.07), and ‘Household items’ (0.48 ± 0.08) layers exhibiting higher reciprocity.

Note, too, that our estimates for mutuality (η_{est}) follow a similar pattern, with the lowest estimates for ‘Advice’, and the highest for ‘Visit’ and ‘Household Items’. These estimates broadly align with theory: reciprocity—and the *expectation* for reciprocity, as represented by the mutuality term—is higher in those relationships that are understood to be more balanced and mutually supportive (i.e., visiting one another’s homes, and borrowing/lending basic household items, like rice and kerosene), and lower in those relationships that are potentially seen as hierarchical and imbalanced (receiving/giving advice, and borrowing/lending money).

We next investigate how reporter ‘reliabilities’ are distributed in these networks. Since the mutuality in these graphs is high ($\eta_{\text{est}} \geq 0.4$), it is very common that reporters repeat the same names across the different name generators. Therefore, it is expected that individual ‘reliability’ terms will play a smaller role in determining the reported social network. Recalling equation (1), this means that the value of θ_m will be small for reporters with a high rate of repeat nominations—that is, the proportion of alters reported by an ego on the ‘give to’ question that gets repeated on the ‘gets from’ question. In the Karnataka dataset, 26% (Advice) to 52% (Household Items) of reporters have an individual rate of repeated nominations of 100%. In such cases, small values of θ_m should not be interpreted as indicating under-reporting, as a small θ_m in this case is just a signal of a high mutuality. The vast majority of reporters (99.49%) with a small θ_m ($\theta_m < 0.1$) in the Karnataka networks have an individual rate of repeat nominations of 100%, regardless of the tie type.

For all four tie types, we observe that reporters tend to under-report relationships, even after having accounted for those individuals who have low θ_m for the reasons discussed above (see Figure S9). This is consistent with prior literature (e.g., Butts, 2003) that suggests that reporters

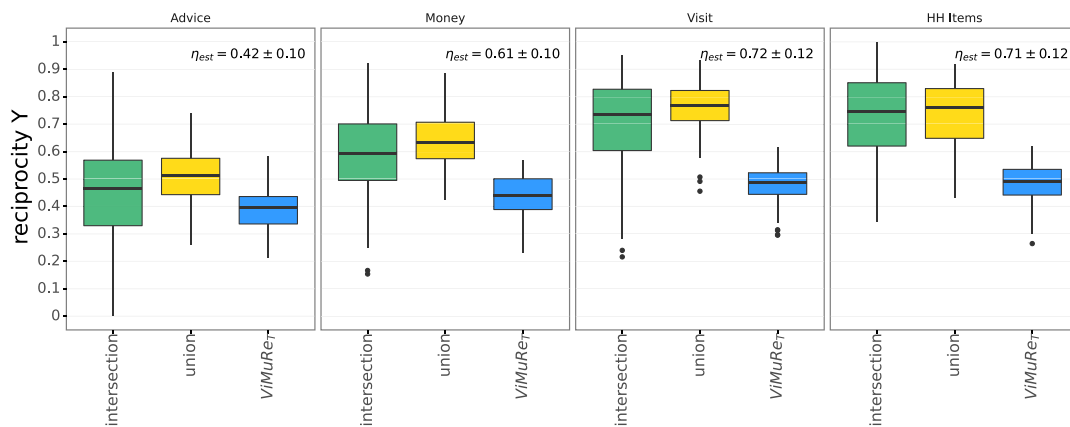


Figure 8. Reciprocity on Karnataka networks. The box plots show the distribution over the 75 networks of the reciprocity measured on the inferred \hat{Y} . Each column is a different tie type, as written on the figure title.

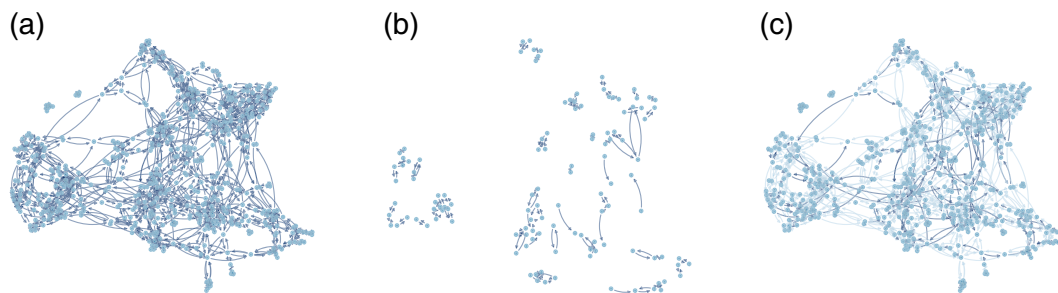


Figure 9. Example of networks estimated by baseline methods and VIMuRe for one Karnataka village (tie type ‘Visit’). (a) Union (recip. = 0.93), (b) intersection (recip. = 0.88), and (c) VIMuRe (recip. = 0.49).

are reasonably accurate when reporting ties, but quite inaccurate when reporting nonties. This effect may be partially induced by the way the questionnaire was formulated, as there were only four entries available for nominating alters. This causes under-reporting to be much more likely, and substantially limits the number of reported tie configurations that can be observed in this system. Indeed, we note a negative correlation between θ_m and the in-degree of ties reported by others involving m (see Figure S11). In particular, reporters nominated by many others (some by more than 20 people) could only nominate up to four among these; such reporters will necessarily have low values for θ_m .

We do not observe any strong differences in the distribution of reliability across the four tie types (see Figure S9). We assess whether reporters are consistent in their reliabilities across the tie types by examining the pairwise distances using the Wasserstein distance (a metric for measuring distances between two distributions; Givens & Shortt, 1984) between each set of tie types (see Figure S10). We see some telling patterns by looking at the consistency of ‘Advice’ with the other tie types: while reporters are most consistent between the ‘Advice’ and ‘Money’ networks, they are least consistent between the ‘Advice’ and ‘Household Items’ or ‘Visit’ networks.

7 Discussion and conclusions

Self-report network data are an important resource for social scientists, but they are also susceptible to several types of reporter bias. Identifying the possible biases that structure self-reports of social relationships remains an essential and open area of research for social network analysis. Failure to identify and account for such reporting biases may lead researchers to draw incorrect inferences (Redhead et al., Accepted). But how should social scientists go about investigating and treating reporting bias in measurements of social networks? We provide a novel statistical solution to—and theoretical and empirical evidence of—this problem, with particular focus on two

forms of bias. First, we investigate and adjust for the general propensity of reporting balanced, reciprocal relationships (i.e., mutuality). Second, we provide a method of accounting for individuals' unique potential to misrepresent or misreport their relationships during reports. Both of these forms of bias have the potential to add substantial 'noise' to empirical representations of social networks, but the extent to which this is present and problematic has not yet been well established. While previous work has explored individual propensities to misreport their ties (e.g., [Butts, 2003](#); [Newman, 2018a](#); [Young et al., 2021](#)), there has been limited formal analysis of the impact that 'mutuality' has on network inference (but see, [Redhead et al., Accepted](#)).

We have focused our attention on cases where multiple reporters are able to provide information on any given relationship. In particular, we have considered 'double-sampled' relationships, where respondents are asked about their role both as giver and as receiver—a common technique that is used in social support network surveys. We have introduced a probabilistic modelling framework, VIMuRe, that provides a principled solution to these issues and aims to more appropriately capture the data generating process associated with name generator designs. VIMuRe takes as input potentially biased, imperfect survey responses and uses these to estimate a 'true' latent network, as well as parameters governing individual biases and relationship-specific tendencies towards mutuality. The model estimates both a ground-truth, Y , and θ which, in certain cases, can be interpreted as a reporter's reliability (conditional on some level of mutuality).

The model that we have introduced here strongly departs from common approaches for dealing with double-sampled network data in the social sciences, in which researchers simply take the union or the intersection of nominations. Our approach also departs from existing network reconstruction methods and advances a framework that is maximally flexible. To our knowledge, existing network reconstruction methods (e.g., [Butts, 2003](#); [Newman, 2018a](#); [Young et al., 2021](#)) that are applicable to social networks focus on the single-sampled case—with the exception of [Redhead et al. \(Accepted\)](#), which is applicable only to double-sampled networks. While we have highlighted double-sampled network data here, our framework can be readily used for many reporting sampling schemes. A tie within a network could be reported on by any number of reporters, up to and including a full 'cognitive social structure' design ([Krackhardt, 1987](#)), where each respondent reports on all other ties in the network. Alongside this, the model remains computationally efficient given the use of variational inference, as opposed to a Monte Carlo approach. Our model can flexibly handle social network datasets of any realistic size, and can scale to large systems of tens of thousands of nodes by exploiting the sparsity of typical network datasets.

Results from our simulation experiments highlight that mutuality dramatically impacts inferred levels of reciprocity. Our results complement previous empirical and theoretical studies (e.g., [Ready & Power, 2021](#); [Redhead et al., Accepted](#)), and show that the simple deterministic approach of taking the union or intersection of nominations leads to biased estimates of reciprocity. Given this, we propose a simple heuristic that is based solely on the mutuality value inferred by the model, that can be used to select the most appropriate point-estimates from an estimated posterior distribution of Y . Findings from our simulation experiments suggest that our approach results in networks that are somewhere between those produced by the union and the intersection. Generally, our approach results in lower levels of reciprocity than deterministic aggregation, because we are appropriately accounting for mutuality.

The importance of considering the core questions of (bias in) network representation—and the utility of VIMuRe—are most clearly demonstrated with our analyses of the empirical data from Karnataka ([Banerjee et al., 2013](#)) and Nicaragua ([Koster, 2018](#)). These datasets result from two very different elicitation approaches, which carry with them different potential risks for bias. The data from Karnataka provide a case where a standard name generator approach was used on a partial sample of the network, and where an upper bound of four was placed on the number of ties that could be reported. This design likely increases the chances that ties are under-reported. In contrast, the data from Nicaragua were collected using a full roster-based design on the entire sample. This approach may inflate the chances that ties are over-reported. In both empirical examples, the prospect of mutuality is salient, as reporters were asked about their roles as givers and receivers in direct succession, and there was no randomization of question order. The results of our empirical applications indicate the importance of mutuality in patterning reports within double-sampled designs. In Karnataka, mutuality values range from ~ 0.4 to ~ 0.7 , and in Nicaragua they are ~ 0.6 . These mutuality values complement the findings from our simulation

experiments, which show that when mutuality is high, taking either the union or the intersection will result in inflated reciprocity values (despite treating discordant responses in very different ways).

Our findings highlight that mutuality is indeed high across a range of different relationship types, and thus the consequences of using these standard deterministic aggregation methods are obvious: a clear disparity between the resulting aggregated networks and the ‘true’ underlying network. The acuteness of this issue depends on the particular tie type, as we can see in the varying levels of mutuality in the Karnataka data (where mutuality is lowest for relationships that may be seen as less balanced). VIMuRe provides a promising way forward here, as it is able to measure and account for mutuality across different sampling regimes.

The empirical examples that we present further elucidate the varying ‘reliabilities’ of reporters—over and above the general propensity to report mutually supportive, reciprocal ties. Importantly, the contrasting results found between the two sets of empirical data reveal general issues with sampling and elicitation, about which practitioners need to be cognisant. The roster-based design used to collect the network data in Nicaragua, resulted in an average of 25 nominations for each prompt. In contrast, given the upper limit of four ties that could be reported in the Karnataka design, the average number of nominations was much lower (around two to three nominations for each prompt) for the various relationship types (see [Table S2](#)). Compounding this issue is the partial sampling procedure implemented in Karnataka. The partial sample included ~46% of the households and ~25% of residents (including children) within the sampled villages. Our findings suggest that when many nominations are of people who were not themselves reporters, there are considerable constraints on the ability to assess reliability. Moreover, our findings suggest that individuals who were named by many others are likely to be seen as ‘unreliable’ (see [Figure S10](#)), in part because these individuals were constrained in their ability to name more than four individuals. Generally speaking, greater coverage of the network and prompts that facilitate collection of more-complete nomination sets will permit more precise estimation of individual ‘reliabilities’ and, thus, more accurate network reconstruction.

Several directions are possible for future improvements to VIMuRe. Our model specifies conditional probabilities, and thus relies on pseudo-likelihood estimation for inferring the parameters. A fruitful avenue for future research is to improve this approximation by characterizing a full joint distribution of a pair of ties ([Contisciani et al., 2022](#)). Doing this may potentially solve the problem of identifying a θ_m for samples with high mutuality and, most importantly, increase the accuracy of estimating posterior distributions for Y . However, any improvement may come at the price of losing analytical tractability, or requiring less flexible approaches. We have focused here on capturing reciprocity, but this does not provide any guarantees of recovering automatically other network properties involving higher-order motifs, such as transitivity or triadic closure (see [Table S2](#)). How to adapt our model to include them is open for future research.

Alongside this, there are several other possibilities for future extensions of the VIMuRe framework that we have introduced here. First, VIMuRe takes as input a set of reported ties and we assumed that this is the only information known. However, if practitioners have access to additional information—such as covariates on nodes—this information could be incorporated into the model. Covariates could also be incorporated into models predicting reliabilities θ , and those reliabilities could vary for senders and receivers as well. For instance, one can consider a suitable prior for the reliabilities θ_m that is based upon a given covariate. It would also be straightforward to extend our model by incorporating more informative priors about the ground-truth network, Y (e.g., if the network had a known block structure). Second, many social networks are fundamentally multi-level, with nodes being nested within higher-order units (e.g., households, businesses, or schools; [Lazega & Snijders, 2015](#)). Formulating an approach to flexibly incorporate multi-level networks further remains an open and important area for extending the VIMuRe framework. Finally, our focus has been on cases where social networks are static. Investigating how to effectively adapt our model for networks evolving over time is an open avenue for future work.

In sum, there are potentially strong biases in self-reported social network data. However, the nature of multiply reported data as containing multiple sources of information about a single underlying relationship permits the application of statistical procedures that can account for such biases. VIMuRe attempts to do this by explicitly modelling mutuality—the tendency of reporters to nominate the same individuals for both directions of a tie—and estimating a reporting

accuracy parameter, θ_m , for each reporter. Model estimation is performed using variational inference, leading to a fast algorithmic implementation that is scalable to large system sizes. Our study of the datasets from Karnataka and Nicaragua establishes that there is indeed important variation in reporters' 'reliability', and that people's reports seem to be driven in part by their normative expectation of relationships as balanced and reciprocal. We observe this high 'mutuality' despite very different data elicitation approaches, and see that it varies based on the type of relationship being elicited. These findings demonstrate the value of employing a tool such as VIMuRe, as it can not only give crucial insights into how social relationships are understood by individuals, but can also provide a way to account for these individual and collective biases and arrive at a more appropriate representation of the network of interest. To facilitate its usage by practitioners, we provide an open source implementation of the code online.

Acknowledgements

This paper comes out of a collaboration funded by a UKRI Economic and Social Research Council Research Methods Development Grant (ES/V006495/1). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Martina Contisciani and Diego Baptista. Caterina De Bacco, Martina Contisciani, and Hadiseh Safdari were supported by the Cyber Valley Research Fund. Daniel Redhead, Cody T. Ross, and Richard McElreath were supported by the Department of Human Behaviour, Ecology, and Culture at the Max Planck Institute for Evolutionary Anthropology.

Data availability

All data and code used in this paper are available in the following public repository: <https://github.com/latentnetworks/vimure>. The data that support the findings of this study are also openly available at the following locations: The Abdul Latif Jameel Poverty Action Lab Dataverse at <https://doi.org/10.7910/DVN/U3BIHX> and The Royal Society Open Science's Electronic [Supplementary Material](https://doi.org/10.1098/rsos.172159) at <https://doi.org/10.1098/rsos.172159>.

Conflict of interest: The authors declare that they have no conflict of interest.

Supplementary material

[Supplementary data](#) are available at *Journal of the Royal Statistical Society* online.

References

- Adams A. M., Madhavan S., & Simon D. (2006). Measuring social networks cross-culturally. *Social Networks*, 28(4), 363–376. <https://doi.org/10.1016/j.socnet.2005.07.007>
- Adams J., & Moody J. (2007). To tell the truth: Measuring concordance in multiply reported network data. *Social Networks*, 29(1), 44–58. <https://doi.org/10.1016/j.socnet.2005.11.009>
- Amini L., Shaikh A., & Schulzrinne H. (2004). Issues with inferring internet topological attributes. *Computer Communications*, 27(6), 557–567. <https://doi.org/10.1016/j.comcom.2003.08.021>
- An W. (2022). You said, they said: A framework on informant accuracy with application to studying self-reports and peer-reports. *Social Networks*, 70, 187–197. <https://doi.org/10.1016/j.socnet.2021.12.006>
- An W., & Schramski S. (2015). Analysis of contested reports in exchange networks based on actors' credibility. *Social Networks*, 40, 25–33. <https://doi.org/10.1016/j.socnet.2014.07.002>
- Ball B., & Newman M. E. (2013). Friendship networks and social status. *Network Science*, 1(1), 16–30. <https://doi.org/10.1017/nws.2012.4>
- Banerjee A., Chandrasekhar A. G., Duflo E., & Jackson M. O. (2013). The diffusion of microfinance. *Science*, 341(6144), 1236498–1–1236498–7. <https://doi.org/10.1126/science.1236498>
- Bell D. C., Belli-McQueen B., & Haider A. (2007). Partner naming and forgetting: Recall of network members. *Social Networks*, 29(2), 279–299. <https://doi.org/10.1016/j.socnet.2006.12.004>
- Bernard H. R., Killworth P., Kronenfeld D., & Sailer L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13(1), 495–517. <https://doi.org/10.1146/annurev.an.13.100184.002431>
- Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–225. <https://doi.org/10.1111/j.2517-6161.1974.tb00999.x>

- Blei D. M., Kucukelbir A., & McAuliffe J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877. <https://doi.org/10.1080/01621459.2017.1285773>
- Borgatti S. P., Mehra A., Brass D. J., & Labianca G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892–895. <https://doi.org/10.1126/science.1165821>
- Brands R. A. (2013). Cognitive social structures in social network research: A review. *Journal of Organizational Behavior*, 34(S1), S82–S103. <https://doi.org/10.1002/job.1890>
- Brewer D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Social Networks*, 22(1), 29–43. [https://doi.org/10.1016/S0378-8733\(99\)00017-9](https://doi.org/10.1016/S0378-8733(99)00017-9)
- Burt R. S. (1984). Network items and the general social survey. *Social Networks*, 6(4), 293–339. [https://doi.org/10.1016/0378-8733\(84\)90007-8](https://doi.org/10.1016/0378-8733(84)90007-8)
- Butts C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks*, 25(2), 103–140. [https://doi.org/10.1016/S0378-8733\(02\)00038-2](https://doi.org/10.1016/S0378-8733(02)00038-2)
- Casella G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2), 83–87. <https://doi.org/10.1080/00031305.1985.10479400>
- Chicco D., & Jurman G. (2020). The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13. <https://doi.org/10.1186/s12864-019-6413-7>
- Contisciani M., Safdari H., & De Bacco C. (2022). Community detection and reciprocity in networks by jointly modeling pairs of edges. *Journal of Complex Networks*, 10(4). <https://doi.org/10.1093/comnet/cnac034>
- De Bacco C., Power E. A., Larremore D. B., & Moore C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4), 042317. <https://doi.org/10.1103/PhysRevE.95.042317>
- D'haeseleer P., & Church G. M. (2004). Estimating and improving protein interaction error rates. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004* (pp. 216–223). IEEE.
- Eagle D. E., & Proeschold-Bell R. J. (2015). Methodological considerations in the use of name generators and interpreters. *Social Networks*, 40, 75–83. <https://doi.org/10.1016/j.socnet.2014.07.005>
- Eagle N., Pentland A. S., & Lazer D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36), 15274–15278. <https://doi.org/10.1073/pnas.0900282106>
- Fehr E., & Gächter S. (1998). Reciprocity and economics: The economic implications of Homo Reciprocans. *European Economic Review*, 42(3–5), 845–859. [https://doi.org/10.1016/S0014-2921\(97\)00131-1](https://doi.org/10.1016/S0014-2921(97)00131-1)
- Freeman L. C. (1992). Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly*, 55(2), 118–127. <https://doi.org/10.2307/2786941>
- Gelman A., Meng X.-L., & Stern H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–760.
- Gelman A., Stern H. S., Carlin J. B., Dunson D. B., Vehtari A., & Rubin D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.
- Givens C. R., & Shortt R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2), 231–240. <https://doi.org/10.1307/mmj/1029003026>
- Heider F. (1958). *The psychology of interpersonal relations*. Wiley.
- Hobson E. A., Silk M. J., Fefferman N. H., Larremore D. B., Rombach P., Shai S., & Pinter-Wollman N. (2021). A guide to choosing and implementing reference models for social network analysis. *Biological Reviews*, 96(6), 2716–2734. <https://doi.org/10.1111/brv.v96.6>
- Hoff P., Fosdick B., Volfovsky A., & Stovel K. (2013). Likelihoods for fixed rank nomination networks. *Network Science*, 1(1), 253–277. <https://doi.org/10.1017/nws.2013.17>
- Hoff P. D., Raftery A. E., & Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460), 1090–1098. <https://doi.org/10.1198/016214502388618906>
- Holland P. W., Laskey K. B., & Leinhardt S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2), 109–137. [https://doi.org/10.1016/0378-8733\(83\)90021-7](https://doi.org/10.1016/0378-8733(83)90021-7)
- Holt-Lunstad J., Smith T. B., Baker M., Harris T., & Stephenson D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, 10(2), 227–237. <https://doi.org/10.1177/1745691614568352>
- Jackson M. O. (2021). Inequality's economic and social roots: The role of social networks and homophily. *SSRN Scholarly Paper ID 3795626*, Social Science Research Network, Rochester, NY.
- Jackson M. O., Rodriguez-Barraquer T., & Tan X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5), 1857–1897. <https://doi.org/10.1257/aer.102.5.1857>
- Karrer B., & Newman M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), 016107. <https://doi.org/10.1103/PhysRevE.83.016107>
- Kenny D. A., & La Voie L. (1984). The social relations model. In *Advances in experimental social psychology* (Vol. 18, pp. 141–182). Academic Press.

- Killworth P., & Bernard H. (1976). Informant accuracy in social network data. *Human Organization*, 35(3), 269–286. <https://doi.org/10.17730/humo.35.3.10215j2m359266n2>
- Koster J. (2018). Family ties: The multilevel effects of households and kinship on the networks of individuals. *Royal Society Open Science*, 5(4), 172159. <https://doi.org/10.1098/rsos.172159>
- Krackhardt D. (1987). Cognitive social structures. *Social Networks*, 9(2), 109–134. [https://doi.org/10.1016/0378-8733\(87\)90009-8](https://doi.org/10.1016/0378-8733(87)90009-8)
- Krackhardt D., & Kilduff M. (1990). Friendship patterns and culture: The control of organizational diversity. *American Anthropologist*, 92(1), 142–154. <https://doi.org/10.1525/aa.1990.92.issue-1>
- Kristiansen S. (2004). Social networks and business success: The role of subcultures in an African context. *American Journal of Economics and Sociology*, 63(5), 1149–1171. <https://doi.org/10.1111/ajes.2004.63.issue-5>
- Lazega E., & Snijders T. A. B. (Eds.) (2015). *Multilevel network analysis for the social sciences: Theory, methods and applications*. Springer.
- Lazer D., Hargittai E., Freelon D., Gonzalez-Bailon S., Munger K., Ognyanova K., & Radford J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866), 189–196. <https://doi.org/10.1038/s41586-021-03660-7>
- Lee F., & Butts C. T. (2018). Mutual assent or unilateral nomination? A performance comparison of intersection and union rules for integrating self-reports of social relationships. *Social Networks*, 55, 55–62. <https://doi.org/10.1016/j.socnet.2018.05.005>
- Lin N. (2002). *Social capital: A theory of social structure and action*. Cambridge University Press.
- Lungeanu A., McKnight M., Negron R., Munar W., Christakis N. A., & Contractor N. S. (2021). Using Trellis software to enhance high-quality large-scale network data collection in the field. *Social Networks*, 66, 171–184. <https://doi.org/10.1016/j.socnet.2021.02.007>
- Marin A. (2004). Are respondents more likely to list alters with certain characteristics? Implications for name generator data. *Social Networks*, 26(4), 289–307. <https://doi.org/10.1016/j.socnet.2004.06.001>
- Marineau J. E., Labianca G. J., Brass D. J., Borgatti S. P., & Vecchi P. (2018). Individuals' power and their social network accuracy: A situated cognition perspective. *Social Networks*, 54, 145–161. <https://doi.org/10.1016/j.socnet.2018.01.006>
- Marsden P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16(1), 435–463. <https://doi.org/10.1146/soc.1990.16.issue-1>
- Marsden P. V. (2003). Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25(1), 1–16. [https://doi.org/10.1016/S0378-8733\(02\)00009-6](https://doi.org/10.1016/S0378-8733(02)00009-6)
- Marsden P. V. (2005). Recent developments in network measurement. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 8–30), No. 27 in Structural Analysis in the Social Sciences. Cambridge University Press.
- Molm L. D. (2010). The structure of reciprocity. *Social Psychology Quarterly*, 73(2), 119–131. <https://doi.org/10.1177/0190272510369079>
- Morris C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American statistical Association*, 78(381), 47–55. <https://doi.org/10.1080/01621459.1983.10477920>
- Newcomb T. M. (1961). The acquaintance process as a prototype of human interaction. In T. M. Newcomb (Ed.), *The acquaintance process* (pp. 259–261). Holt, Rinehart & Winston.
- Newman M. E. J. (2018a). Estimating network structure from unreliable measurements. *Physical Review E*, 98(6), 062321. <https://doi.org/10.1103/PhysRevE.98.062321>
- Newman M. E. J. (2018b). Network structure from rich but noisy data. *Nature Physics*, 14(6), 542–545. <https://doi.org/10.1038/s41567-018-0076-1>
- Nolin D. A. (2008). *Food-sharing networks in Lamalera, Indonesia: Tests of adaptive hypotheses* [Ph.D. thesis]. University of Washington.
- Nolin D. A. (2010). Food-sharing networks in Lamalera, Indonesia: Reciprocity, kinship, and distance. *Human Nature*, 21(3), 243–268. <https://doi.org/10.1007/s12110-010-9091-3>
- Park P. S., Blumenstock J. E., & Macy M. W. (2018). The strength of long-range ties in population-scale social networks. *Science*, 362(6421), 1410–1413. <https://doi.org/10.1126/science.aau9735>
- Peixoto T. P. (2018). Reconstructing networks with unknown and heterogeneous errors. *Physical Review X*, 8(4), 041011. <https://doi.org/10.1103/PhysRevX.8.041011>
- Perkins J. M., Subramanian S. V., & Christakis N. A. (2015). Social networks and health: A systematic review of sociocentric network studies in low- and middle-income countries. *Social Science & Medicine*, 125, 60–78. <https://doi.org/10.1016/j.socscimed.2014.08.019>
- Power E. A. (2017). Social support networks and religiosity in rural South India. *Nature Human Behaviour*, 1(3), 0057. <https://doi.org/10.1038/s41562-017-0057>
- Pustejovsky J. E., & Spillane J. P. (2009). Question-order effects in social network name generators. *Social Networks*, 31(4), 221–229. <https://doi.org/10.1016/j.socnet.2009.06.001>

- Ready E., & Power E. A. (2018). Why wage earners hunt: Food sharing, social structure, and influence in an Arctic mixed economy. *Current Anthropology*, 59(1), 74–97. <https://doi.org/10.1086/696018>
- Ready E., & Power E. A. (2021). Measuring reciprocity: Double sampling, concordance, and network construction. *Network Science*, 9(4), 387–402. <https://doi.org/10.1017/nws.2021.18>
- Redhead D., McElreath R., & Ross C. T. (Accepted). Reliable network inference from unreliable data: A tutorial on latent network modeling using STRAND. *Psychological Methods*. Advance online publication. <https://psyarxiv.com/mkp2y/>
- Redhead D., & Power E. A. (2022). Social hierarchies and social networks in humans. *Philosophical Transaction of the Royal Society B*, 377(1845), 20200440. <https://doi.org/10.1098/rstb.2020.0440>
- Redhead D., & von Rueden C. R. (2021). Coalitions and conflict: A longitudinal analysis of men's politics. *Evolutionary Human Sciences*, 3, E31. <https://doi.org/10.1017/ehs.2021.26>
- Robbins H. E. (1955). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Probability and Statistics* (pp. 157–164).
- Robins G., Pattison P., Kalish Y., & Lusher D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29(2), 173–191. <https://doi.org/10.1016/j.socnet.2006.08.002>
- Ross C. T., & Redhead D. (2022). DieTryin: An R package for data collection, automated data entry, and post-processing of network-structured economic games, social networks, and other roster-based dyadic data. *Behavior Research Methods*, 54(2), 611–631. <https://doi.org/10.3758/s13428-021-01606-5>
- Safdari H., Contisciani M., & De Bacco C. (2021). Generative model for reciprocity and community detection in networks. *Physical Review Research*, 3(2), 023209. <https://doi.org/10.1103/PhysRevResearch.3.023209>
- Sewell D. K. (2019). Latent space models for network perception data. *Network Science*, 7(2), 160–179. <https://doi.org/10.1017/nws.2019.1>
- Shakya H. B., Christakis N. A., & Fowler J. H. (2017). An exploratory comparison of name generator content: Data from rural India. *Social Networks*, 48, 157–168. <https://doi.org/10.1016/j.socnet.2016.08.008>
- Simpson B., Markovsky B., & Steketee M. (2011). Power and the perception of social networks. *Social Networks*, 33(2), 166–171. <https://doi.org/10.1016/j.socnet.2010.10.007>
- Simpson C. R. (2022). On the structural equivalence of coresidents and the measurement of village social structure. *Social Networks*, 69, 55–73. <https://doi.org/10.1016/j.socnet.2020.02.010>
- Smith K. P., & Christakis N. A. (2008). Social networks and health. *Annual Review of Sociology*, 34(1), 405–429. <https://doi.org/10.1146/soc.2008.34.issue-1>
- Sosa J., & Rodríguez A. (2021). A latent space model for cognitive social structures data. *Social Networks*, 65, 85–97. <https://doi.org/10.1016/j.socnet.2020.12.002>
- Sprinzak E., Sattath S., & Margalit H. (2003). How reliable are experimental protein–protein interaction data? *Journal of Molecular Biology*, 327(5), 919–923. [https://doi.org/10.1016/S0022-2836\(03\)00239-0](https://doi.org/10.1016/S0022-2836(03)00239-0)
- Titterton D., Smith A., & Makov U. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
- von Rueden C., Redhead D., O’Gorman R., Kaplan H., & Gurven M. (2019). The dynamics of men’s cooperation and social status in a small-scale society. *Proceedings of the Royal Society B: Biological Sciences*, 286(1908), 20191367. <https://doi.org/10.1098/rspb.2019.1367>
- Warner R. M., Kenny D. A., & Stoto M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, 37(10), 1742–1757. <https://doi.org/10.1037/0022-3514.37.10.1742>
- Young J.-G., Cantwell G. T., & Newman M. E. J. (2021). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 8(6), cnaa046. <https://doi.org/10.1093/comnet/cnaa046>
- Yousefi-Nooraie R., Marin A., Hanneman R., Pullenayegum E., Lohfeld L., & Dobbins M. (2019). The relationship between the position of name generator questions and responsiveness in multiple name generator surveys. *Sociological Methods & Research*, 48(2), 243–262. <https://doi.org/10.1177/0049124117701484>

Inference of hyperedges and overlapping communities in hypergraphs

Received: 18 May 2022

Martina Contisciani¹✉, Federico Battiston² & Caterina De Bacco¹✉

Accepted: 2 November 2022

Published online: 24 November 2022

 Check for updates

Hypergraphs, encoding structured interactions among any number of system units, have recently proven a successful tool to describe many real-world biological and social networks. Here we propose a framework based on statistical inference to characterize the structural organization of hypergraphs. The method allows to infer missing hyperedges of any size in a principled way, and to jointly detect overlapping communities in presence of higher-order interactions. Furthermore, our model has an efficient numerical implementation, and it runs faster than dyadic algorithms on pairwise records projected from higher-order data. We apply our method to a variety of real-world systems, showing strong performance in hyperedge prediction tasks, detecting communities well aligned with the information carried by interactions, and robustness against addition of noisy hyperedges. Our approach illustrates the fundamental advantages of a hypergraph probabilistic model when modeling relational systems with higher-order interactions.

Over the past twenty years, networks have allowed to map and characterize the architecture of a wide variety of relational data, from social and technological systems to the human brain¹. Despite their success, traditional graph representation are unable to provide a faithful representation of the patterns of interactions occurring in the real-world². Collections of nodes and links—networks—can only properly encode dyadic relations. Yet, in the last few years systems as diverse as cellular networks³, structural and functional brain networks^{4,5}, social systems⁶, ecosystems⁷, social image search engines⁸, human face-to-face interactions⁹ and collaboration networks¹⁰, have shown that a large fraction of interactions occurs among three or more nodes at a time. These higher-order systems are hence best described by different mathematical frameworks such as hypergraphs¹¹, where hyperedges of arbitrary dimensions may encode structured relations among any number of system units^{12–14}. Interestingly, providing a higher-order description of the system interactions has been shown to lead to the emergence of new collective phenomena¹⁵ in diffusive^{16,17}, synchronization^{18–22}, spreading^{23–25}, and evolutionary²⁶ processes.

To properly describe the higher-order organization of real-world networks, a variety of growing^{27,28} and equilibrium models, such as generalized configuration models^{29–31} have been proposed. Tools from

topological data analysis have allowed to obtain insights into the higher-order organization of real-world networks^{32,33}, and methods to infer higher-order interactions from pairwise records have been suggested³⁴. Finally, several powerful network metrics and ideas have been extended beyond the pair, from higher-order clustering³⁵, spectral methods³⁶ and centrality^{37,38} to motifs³⁹ and network backbone⁴⁰.

Despite a few recent contributions^{41–47}, how to define and identify the mesoscale organization of real-world hypergraphs is still a largely unexplored topic. Here, we propose a new principled method to extract higher-order communities based on statistical inference. More broadly, our approach is that of generative models, which incorporate a priori community structure by means of latent variables, inferred directly from the observed interactions^{48–50}. Beyond its efficient numerical implementation, our model has several desirable features. It detects overlapping communities, an aspect that is missing in current approaches of community detection in hypergraphs and that is arguably better representative of scenarios where nodes are expected to belong to multiple groups. It also provides a natural measure to perform link prediction tasks, as it outputs the probability that a given hyperedge exists between any subset of nodes. Similarly, it allows to generate synthetic hypergraphs with given community structure, an ingredient that can be given in input or learned from data. Moreover,

¹Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany. ²Department of Network and Data Science, Central European University, 1100 Vienna, Austria. ✉ e-mail: martina.contisciani@tuebingen.mpg.de; caterina.debacco@tuebingen.mpg.de

our explicit higher-order approach is not only more grounded theoretically, but also more efficient than applying graph algorithms to higher-order data projected into pairwise records.

We apply our method to a variety of real-world systems, showing that it recovers communities more robustly against noisy addition of large hyperedges than methods on projected pairwise data, it achieves high performance in predicting missing hyperedges, and it allows to determine the influence of hyperedge size in such prediction tasks. We also illustrate how our higher-order approach detects communities that are more aligned with the information carried by hyperedges than what is recorded by node attributes. Through these examples, we illustrate how a principled higher-order probabilistic approach can shed light on the role that higher-order interactions play in real-world complex systems.

Results

The Hypergraph-MT model

Here, we introduce Hypergraph-MT, a probabilistic generative model for hypergraphs with mixed-membership community structure. Based on a statistical inference framework, our model provides a principled, efficient and scalable approach to extract overlapping communities in networked systems characterized by the presence of interactions beyond the pair.

At its core, our approach assumes that nodes belong to different groups in different amounts, as specified by a set of membership vectors. These memberships then determine the probability that any subset of nodes is connected with a hyperedge. We denote a hypergraph with N nodes $\mathcal{V} = \{i_1, \dots, i_N\}$ and E hyperedges $\mathcal{E} = \{e_1, \dots, e_E\}$ as $\mathcal{H}(\mathcal{V}, \mathcal{E})$. Mathematically, this can be represented as an adjacency tensor A with entries A_{i_1, \dots, i_d} equal to the weight of a d -dimensional interaction between the nodes i_1, \dots, i_d . For instance, for contact interactions, A_{i_1, \dots, i_d} could be the number of times that nodes i_1, \dots, i_d were in close contact together.

Given these definitions, we can specify the likelihood of observing the hypergraph given a set of latent variables θ , which include the membership vectors. This relies on modeling $P(A_{i_1, \dots, i_d} | \theta)$, the probability of observing a hyperedge given θ . We model this probability as:

$$P(A_{i_1, \dots, i_d} | \theta) = \text{Pois}(A_{i_1, \dots, i_d}; \lambda_{i_1, \dots, i_d}), \tag{1}$$

where $\lambda_{i_1, \dots, i_d} = \sum_{k_1, \dots, k_d} u_{i_1 k_1} \dots u_{i_d k_d} w_{k_1, \dots, k_d}$. The set of latent variables is defined by $\theta = (\mathbf{u}, \mathbf{w})$, where \mathbf{u} is a $N \times K$ -dimensional community membership matrix and \mathbf{w} is an affinity tensor, which captures the idea that an interaction is more likely to exist between nodes of compatible communities. If only pairwise interactions exist, the affinity matrix has dimension $K \times K$. Therefore, the problem reduces to the traditional network case and can be efficiently solved⁴⁹. When higher-order interactions are present, the dimension of the affinity tensor \mathbf{w} can become arbitrarily large depending on the size d_e of a hyperedge e , i.e., the number of nodes present in it. In fact, \mathbf{w} has as many entries as all the possible d_e -way interactions between all K groups. For instance, in a hypergraph with only 2-way and 3-way interactions, we have $\mathbf{w} = [\mathbf{w}^{(2)}, \mathbf{w}^{(3)}]$ with $\mathbf{w}^{(2)}$ of dimension $K \times K$ and $\mathbf{w}^{(3)}$ of dimension $K \times K \times K$.

The question is thus how to reduce the dimension of \mathbf{w} . A relevant choice that overcomes these problems is that of assortativity⁵¹, implying that a hyperedge is more likely to exist when all nodes in it belong to the same group. This captures well situations where homophily, the tendency of nodes with similar features to be connected to each other, plays a role, as observed in social or biological networks^{49,52}. Mathematically, the only non-zero elements of \mathbf{w} are the “diagonal” ones, that is:

$$w_{k_1, \dots, k_d} = \delta_{k_1, \dots, k_d} w_{k_1, \dots, k_d}. \tag{2}$$

With this, we obtain a matrix \mathbf{w} of dimension $D \times K$, where $D = \max_{e \in \mathcal{E}} d_e$ is the maximum hyperedge size in the dataset. In principle, one could envisage other ways to restrict \mathbf{w} to control its dimension. However, we found that the choice in Eq. (2) provides a natural interpretation, results in good prediction performance on both real and synthetic datasets, and is computationally scalable. A similar problem of dimensionality reduction has been tackled in ref. 45, which investigated the more constrained case of hard-membership models.

Putting all together, we model the likelihood of the hypergraph as:

$$P(\mathbf{A} | \theta) = \prod_{e \in \Omega} e^{-\lambda_e} \frac{\lambda_e^{A_e}}{A_e!}, \tag{3}$$

$$\text{with } \lambda_e = \sum_k w_{d_e k} \prod_{i \in e} u_{ik}, \tag{4}$$

where $\Omega = \{e | e \subseteq \mathcal{V}, d_e \geq 2\}$ is the set of all potential hyperedges. In practice, we can reduce this space by considering only the possible hyperedges of a certain size lower or equal than the maximum observed size D . In Eq. (3) we assumed conditional independence between hyperedges given the latent variables, a standard assumption in these types of models. Such a condition could in principle be relaxed following the approaches of refs. 53–55, we do not explore this here.

Having defined Eq. (3), the goal is to infer the latent variables \mathbf{u} and \mathbf{w} given the observed hypergraph \mathbf{A} . To infer the values of $\theta = (\mathbf{u}, \mathbf{w})$, we consider both maximum likelihood estimation (assuming uniform priors on the parameters) and maximum a posteriori estimation (assuming non-uniform priors). The derivations are similar and rely on an efficient expectation-maximization (EM) algorithm⁵⁶ that exploits the sparsity of the dataset, as detailed in the Methods section and in the Supplementary Note 1.

We obtain the following algorithmic updates for the membership vectors:

$$u_{ik} = \frac{\sum_{e \in \mathcal{E}} B_{ie} \rho_{ek}}{\sum_{e \in \Omega | i \in e} w_{d_e k} \prod_{j \in e | j \neq i} u_{jk}}, \tag{5}$$

where B_{ie} is equal to the weight of the hyperedge e to which the node i belongs (it is an entry of the hypergraph incidence matrix) and ρ is a variational distribution determined in the expectation step of the EM procedure. The numerator of Eq. (5) can be computed efficiently, as we only need the non-zero entries of the incidence matrix, which is typically sparse. Instead, computing the denominator can be prohibitive depending on the value of D , the maximum hyperedge size. This is due to the summation over all possible hyperedges in Ω , which requires extracting all possible combinations $\binom{N}{d}$, for $d = 2, \dots, D$.

This problem is not present in the case of graphs, as this summation would be over N^2 terms at most. This issue clearly highlights the importance of algorithmic efficiency in handling hypergraph data, an aspect that cannot be overlooked to make a model work in practice.

We propose a solution to this problem that reduces the computational complexity to $O(NDK)$ and makes our algorithm efficient, scalable and applicable in practice. The key is to rewrite the summation over Ω such that we have an initial value that can be updated at cost $O(1)$ after each update $u_{ik}^{(t)} \rightarrow u_{ik}^{(t+1)}$, which can be done in parallel over $k = 1, \dots, K$. This formulation is explained in details in the Supplementary Note 1, where we also show how to edit the updates in Eq. (5) by imposing sparsity (with a proper prior distribution) or by constraining the membership vectors to be probability vectors such that $\sum_k u_{ik} = 1$. In both cases, we get a constant term added in the denominators of the updates.

Table 1 | Summary of higher-order datasets

	N	E	E_G	M	M_G	$\langle k \rangle$	$s(k)$	$\langle d \rangle$	$s(d)$	D	% $d=2$	% $d > 2 \in G$	K
High school	327	7818	5818	172,035	189,928	55.6	27.1	2.3	0.5	5	70.3%	88.5%	9
Primary school	242	12,704	8317	106,879	127,886	127.0	55.2	2.4	0.6	5	61.0%	87.5%	11
Workplace	92	788	755	9645	9831	17.7	8.6	2.1	0.3	4	94.2%	88.2%	5
Hospital	75	1825	1139	27,835	32,788	59.1	49.0	2.4	0.6	5	60.7%	95.1%	4
Gene-Disease	4642	2738	55,795	4131	114,444	1.7	3.6	5.8	5.2	25	32.4%	0.6%	25
Justice	38	2826	264	15,040	190,790	366.7	203.6	4.9	1.7	9	7.6%	81.8%	2
House bills	1494	41,362	360,086	47,212	2,451,751	245.8	251.6	8.9	6.6	24	18.5%	2.1%	2
Senate bills	293	19,872	22,157	27,300	732,561	482.0	396.9	7.1	5.4	24	16.5%	14.8%	2
House committees	289	106	2535	111	4312	0.7	2.0	8.6	3.6	18	0.9%	0.0%	2
Senate committees	282	275	12,761	289	41,008	16.2	12.6	16.6	6.0	25	0.0%	0.0%	2
Walmart	1025	3553	8029	5112	13,769	9.8	16.7	2.8	1.2	11	51.0%	7.0%	10
Trivago	6687	33,963	69,875	40,280	115,533	13.9	13.8	2.7	1.3	26	59.6%	16.1%	36

Shown are the number of nodes (N), number of hyperedges in the hypergraph (E) and in the graph (E_G), number of weighted hyperedges in the hypergraph (M) and in the graph (M_G), mean node degree ($\langle k \rangle$), SD of node degree ($s(k)$), mean hyperedge size ($\langle d \rangle$), SD of hyperedge size ($s(d)$), maximum hyperedge size (D), percentage of pairwise interactions (% $d=2$), percentage of pairwise interactions in the 2-combination set of hyperedges of size bigger than 2 that are already in the graph (% $d > 2 \in G$), and number of communities (K).

Finally, the updates of the affinity matrix are given by:

$$w_{dk} = \frac{\sum_{e \in \mathcal{E} | d_e = d} A_e \rho_{ek}}{\sum_{e \in \mathcal{E} | d_e = d} \prod_{j \in e} u_{jk}} \quad (6)$$

These are also computationally efficient to implement and can be updated in parallel. Further details are in the Methods section and in the Supplementary Note 1, where we also provide a pseudocode for the whole inference routine. Additionally, in the Supplementary Note 2 we show the validation of our model on synthetic data with ground-truth community structure and the comparison against the generative method of ref. 45 and the spectral method of ref. 47. Hypergraph-MT shows a strong and increasing performance in recovering communities as the ground-truth community structure becomes stronger, similarly to the method of ref. 47. However, this method is designed to capture hard-membership communities and benefits from having an inference routine similar to the generative process of the synthetic data. In particular, Hypergraph-MT significantly outperforms the competing methods Graph-MT, Pairs-MT, and that of ref. 45 as soon as the ground-truth community structure becomes less noisy. Remarkably, this is observed in synthetic datasets that are generated with a different generating process than that of Hypergraph-MT. As a consequence, the positive performance of our method confirms the robustness and the reliability of the methodology here introduced.

Results on empirical data

We analyze hypergraphs derived from empirical data from various domains. For each one, we report a diverse range of structural properties such as number of nodes, hyperedges and their sizes, as detailed in Table 1. Moreover, the datasets provide node metadata, which we use to fix the number of communities K , aiming to compare the resulting communities with this additional information. For further details on the datasets, see the Methods section. For each hypergraph, we run Hypergraph-MT ten times with different random initialization and select the result with the highest likelihood. For comparison, we run the model on two baselines structures obtained from the same empirical data: a graph obtained from clique expansions of each hyperedge (Graph-MT), where a hyperedge of size d is decomposed in $\frac{d(d-1)}{2}$ unordered pairwise interactions; a graph obtained using only hyperedges with $d_e = 2$ (Pairs-MT). Notice that running our model on graphs reduces to MULTITENSOR—the model presented in ref. 49—with an assortative affinity matrix. As a remark, we use interchangeably the terms *graph* or *network* to refer to the data with only pairwise interactions, and the term *hypergraph* for the higher-order data.

The advantage of using hypergraphs. The goal of using the two baselines is to assess the advantage (if any) in treating a dataset with higher-order interactions as a hypergraph. Indeed, in practice higher-order data are often reduced to their projected graph, an operation, which not only generates a potentially misleading loss of information, but which is also computationally expensive⁴¹. Hence, before evaluating the performance of Hypergraph-MT on various datasets, we turn to the following fundamental question: given a dataset of high-order interactions, does a hypergraph representation bring any advantage compared to a simpler graph representation? If the answer is positive, then we should analyze the data with an algorithm that handles hypergraphs. If not, a simpler network algorithm should be enough.

To this end, we analyze four datasets describing human close-proximity contact interactions obtained from wearable sensor data at a high school (High school), a primary school (Primary school), a workplace (Workplace) and a hospital (Hospital). For the analysis, we run the model on the three different structures (hypergraph, clique expansions, and pairwise edges) described above. For each dataset, we compare the inferred partitions with the node metadata that describe either the classes, the departments, or the roles the nodes belong to. We measure closeness to the metadata with the F1-score, a measure for hard-membership classification. It ranges between 0 and 1, where 1 indicates perfect matching between inferred and given partitions. Table 2 shows the performance with the different structures, and both hypergraphs and graphs perform similarly. Notice that the average size of hyperedges in these datasets is around 2.2; thus interactions are mainly pairwise to start with. Moreover, interactions with $d_e > 2$ include people who already interact pairwise (see column % $d > 2 \in G$ in Table 1). Hence, a clique expansion of these is not expected to provide much distinct information from that already present in the pairwise subset of the dataset. Overall, these results suggest that hypergraphs do not bring any additional advantage for these types of datasets, and running a network algorithm would be enough.

To understand how this assessment may change, we present a toy example built from the High school dataset. We select the subset of nodes belonging to two classes (2BIO1 and MP*2 in our example), and we manipulate it by artificially adding a large hyperedge. It simulates an event where ten external people (guests) and a random subset of ten existing nodes are participating. This is represented by the gray hyperedge of dimension 20 in Fig. 1 (left). Here, the green nodes are the external guests, while the blue and orange nodes are the randomly-selected students from the two classes, respectively. While we only add one hyperedge, its size significantly differs from that of all the other existing hyperedges. In

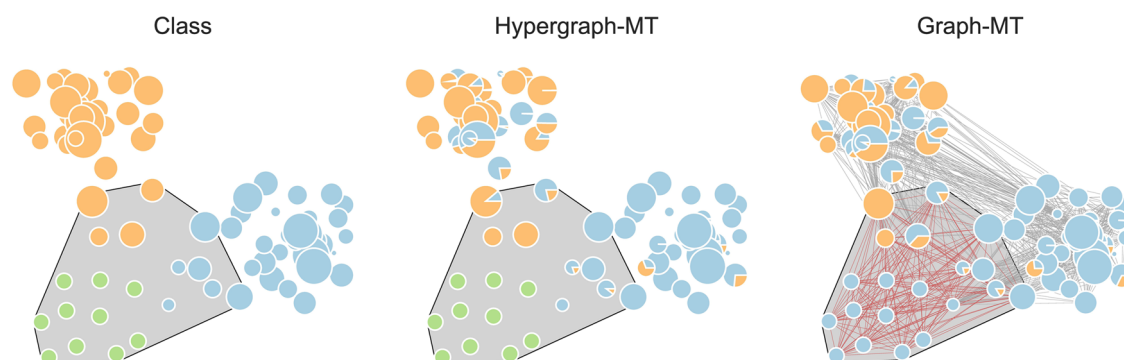


Fig. 1 | The advantage of hypergraph representation: an illustrative example.

The left plot shows a subset of the High school dataset, with nodes belonging to the classes 2BIO1 (light blue) and MP*2 (orange), and ten external guests (green). Node size is proportional to the degree. The gray hyperedge simulates an event, and we omit the other hyperedges for visualization clarity. The central plot displays the partition extracted by Hypergraph-MT and on the right we find the partition

extracted by Graph-MT. In the latter, the gray edges denote the interactions in the graph (obtained by clique expansions) before the event, and the red edges are the interactions added because of the simulated event. This example shows the advantage of using hypergraphs as this representation is more resilient to the addition of a noisy hyperedge and is more robust in detecting communities.

particular, a clique expansion resulting from this additional hyperedge brings in $\binom{20}{2}$ new edges of size 2 (red in the figure). Hence, we expect this additional information to impact the structure of Graph-MT much more than the hypergraph. Figure 1 shows that Hypergraph-MT is not biased by the presence of this individual large hyperedge, and it well recovers the external guests by assigning zero memberships to them for both classes. Conversely, Graph-MT assigns the guests to the blue class. With this toy example, we show a possible scenario where hypergraphs have an advantage, as this representation is more resilient to the addition of a noisy hyperedge and is more robust in detecting communities.

Hyperedge prediction: analysis of a Gene-Disease dataset. We now turn our attention to the analysis of a higher-order Gene-Disease dataset, where nodes are genes, and a hyperedge connects genes that are associated with a disease. Here, we focus on the ability of our model to predict missing hyperedges. We measure prediction performance using a cross-validation protocol where hyperedges are divided into train and test sets. The train set is used for parameter estimation, while performance is evaluated on the test set. We compute the area under the receiver-operator curve (AUC), and use the probability assigned by our model of a hyperedge to exist as input scores for this metric. For Graph-MT, the probability of a hyperedge to exist is computed as the product of the probabilities that each single edge exists. For details, see the Methods section. When evaluating Pairs-MT, we measure the AUC on the subset of test hyperedges of size 2. To perform a balanced comparison in this case, we also measure the AUC for both Hypergraph-MT and Graph-MT on this set (pairs), while still training on the whole train set. This provides information on the utility of large hyperedges to predict pairwise interactions.

Table 2 | Comparison of community detection algorithms in human close-proximity contact interactions datasets

	Hypergraph-MT	Graph-MT	Pairs-MT
High school	0.757	0.776	0.755
Primary school	0.907	0.916	0.928
Workplace	0.829	0.820	0.830
Hospital	0.580	0.491	0.554

For each dataset, we show the F1-score obtained by comparing a node metadata against the inferred partitions from the hypergraphs (Hypergraph-MT), the graphs obtained by clique expansions (Graph-MT), and the graphs given only by the registered pairwise interactions (Pairs-MT).

We vary the maximum hyperedge size D to show how each method responds to the incorporation of progressively larger edges in terms of prediction tasks. Interestingly, we observe a strong shift in performance around $D=15,16$, where Hypergraph-MT significantly outperforms Graph-MT and Pairs-MT (see Fig. 2). This highlights that hyperedges with larger size carry useful information that cannot be fully captured via clique expansions. This is true regardless of the type of missing edges being predicted (hyperedges or pairs-only). In addition, predictive performance is improved homogeneously across hyperedge sizes in the held-out set. Namely, we are not improving just in predicting the pairs-only, as shown by Hypergraph-MT (pairs), but also those of bigger sizes, see Supplementary Fig. 3. This is where Graph-MT fails because the additional information introduced by the clique expansions produces a much denser graph than the input data that may not be correlated with the true existing hyperedges, thus blurring the observations given in the input. These results not only highlight the ability of our model to predict missing data, but also how the knowledge of large hyperedges helps the prediction of hyperedges of smaller sizes.

Overlapping communities and interpretability: analysis of a Justice dataset. Together with hyperedge prediction, Hypergraph-MT allows to extract relevant information also on the mesoscale organization of real-world hypergraphs. As a case study, we analyze a dataset recording all the votes expressed by the Justices of the Supreme Court in the U.S. from 1946 to 2019 case by case. Justices are nodes, and hyperedges connect Justices that expressed the same vote in a given case. The structure of this hypergraph is different from the others analyzed above: it has fewer nodes ($N=38$) but it is denser ($E=2826$), on average a Justice votes 367 times. Similarly, the graph obtained with clique expansion has substantially fewer edges ($E_G=264$) but with higher weights than the hypergraph. See Table 1 for details. Examining the communities inferred in these two markedly distinct structures can provide direct insights into the particular aspects captured by a hypergraph formulation. To this end, we compare the inferred partitions with the political parties of the Justices, i.e., Democrat or Republican, information provided as node metadata. We use the cosine similarity (CS), a metric that measures the distance between vectors, and thus it is better suited to capture mixed-membership communities. The CS varies between 0 and 1, where 1 means that the inferred partition matches perfectly the one shown by political affiliation. For each node, we compute the CS between its political party and the partitions inferred by Hypergraph-MT and Graph-MT. Figure 3a shows the point-by-point comparison between the resulted cosine similarities of the two methods. Here, each marker is a Justice

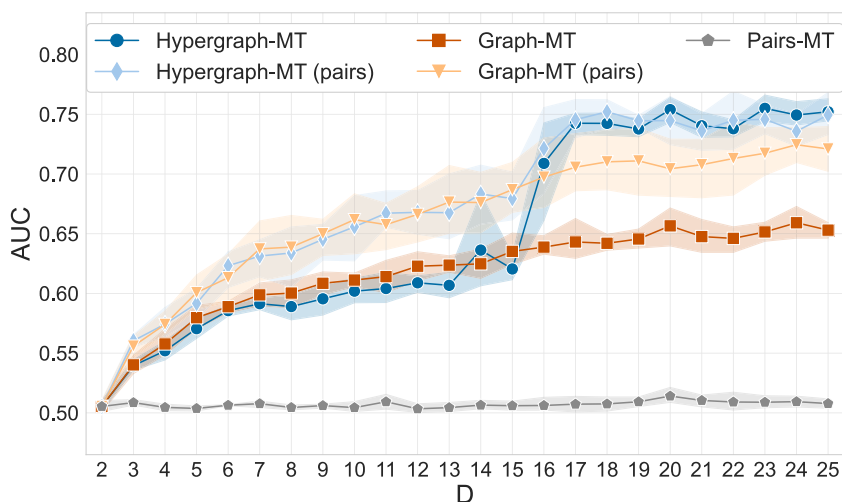


Fig. 2 | Critical size for hyperedge prediction in a Gene-Disease dataset. We measure the AUC by varying the maximum hyperedge size D . The results are averages and standard deviations over 5-fold cross-validation test sets, and the baseline for AUC is the random value 0.5. We run the model on the hypergraphs (Hypergraph-MT), the graphs obtained by clique expansions (Graph-MT), and the graphs given only by the registered pairwise interactions (Pairs-MT). To perform a

balanced comparison against Pairs-MT, for Hypergraph-MT and Graph-MT we additionally measure the AUC on the subset of test hyperedges of size 2 (pairs), while still training on the whole train set. The plot shows the existence of a critical hyperedge size beyond which the higher-order algorithm significantly outperforms alternative methods.

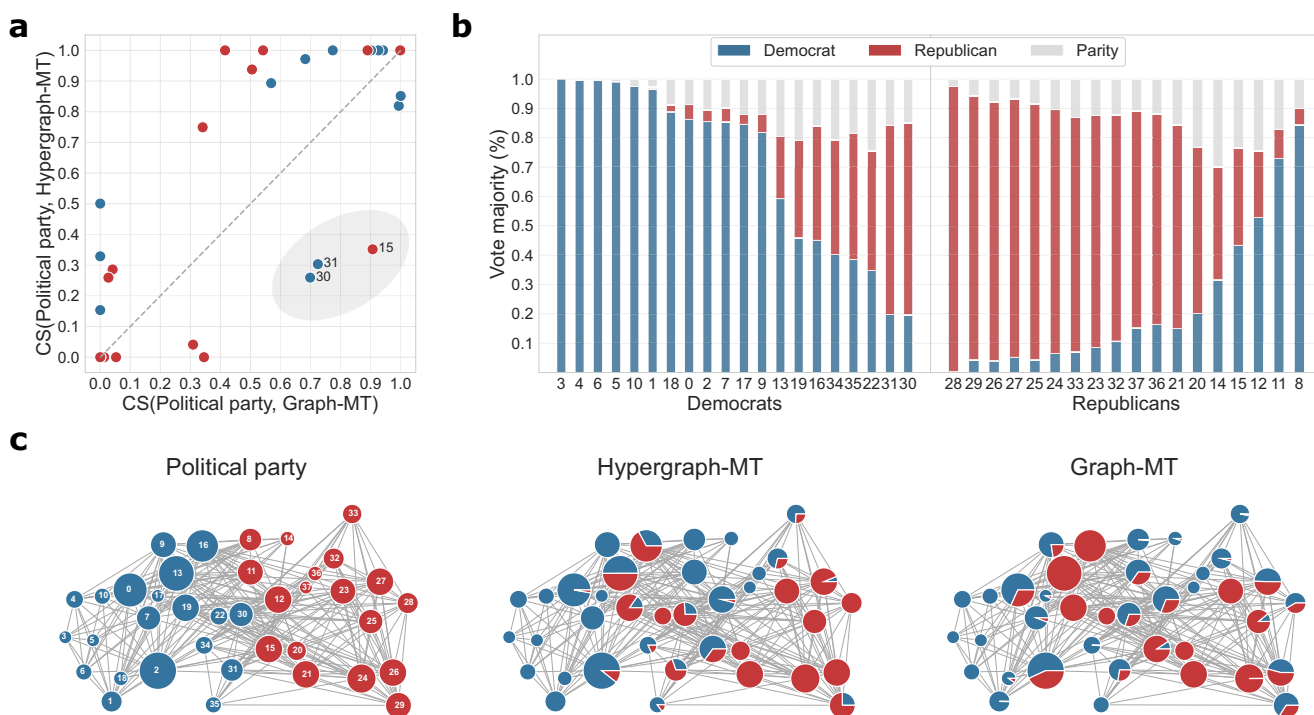


Fig. 3 | Inference of overlapping communities in a co-voting higher-order dataset of the U.S. Justices. **a** Point-by-point comparison between the cosine similarities (CS) obtained by Hypergraph-MT and Graph-MT. For each Justice (marker in the plot), we compute the CS between the partitions inferred by the methods and the political party of the Justices, i.e., Democrat (blue) and Republican (red). **b** Vote majority proportion of the hyperedges of each Justice. Every hyperedge is colored based on the majority political party of the Justices involved in it,

i.e., either Democratic, Republican, or equally distributed (gray). Then, for every Justice, we extract the percentage of times that they participate in hyperedges of a given majority. **c** Data partition according to the political party (left), and the mixed-membership communities inferred by Hypergraph-MT (center) and Graph-MT (right). Node size is proportional to the degree, node labels are Justice IDs, and the interactions are the edges of the projected graph.

and colors represent their political parties. Points above (below) the diagonal represent Justices for which the communities inferred by Hypergraph-MT (Graph-MT) align better with the political party. In several cases the two models infer memberships that align similarly with political affiliation: upper-right corner, where both models are

aligned well, and lower-left corner, where they are both not aligned well. The interesting behavior is shown in the bottom-right area highlighted in gray, containing three Justices whose political affiliations are more closely associated with the communities inferred by Graph-MT than those of Hypergraph-MT. To investigate these cases,

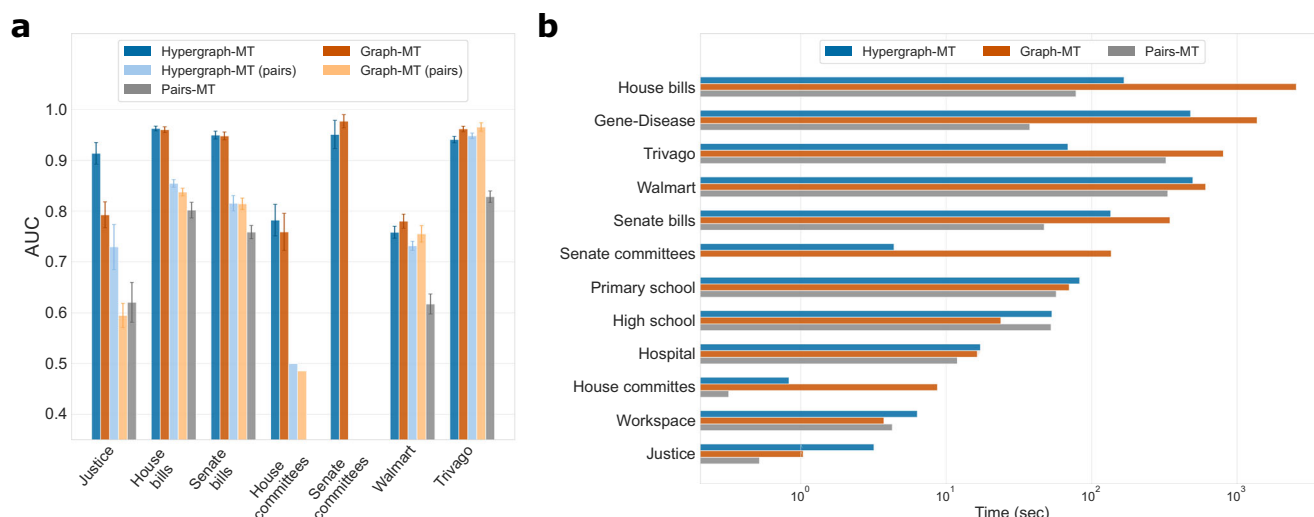


Fig. 4 | Hyperedge prediction performance and computational complexity in higher-order datasets. **a** The performance of hyperedge prediction is measured with the AUC, whose baseline is the random value 0.5. The results are averages and standard deviations over 5-fold cross-validation test sets. For each dataset, we run the model on the hypergraphs (Hypergraph-MT), the graphs obtained by clique expansions (Graph-MT), and the graphs given only by the registered pairwise

interactions (Pairs-MT). To perform a balanced comparison against Pairs-MT, for Hypergraph-MT and Graph-MT we additionally measure the AUC on the subset of test hyperedges of degree 2 (pairs), while still training on the whole train set. **b** Computational complexity of Hypergraph-MT, Graph-MT, and Pairs-MT for the different higher-order datasets. We show the running time for one realization.

we inspect the information carried by the hyperedges. Specifically, for each hyperedge we measure the majority political party based on the affiliation of the Justices involved in it. For instance, a hyperedge of size 5 made of 4 democrats and 1 republican has a Democratic majority. We also account for ties, when equal numbers of Justices are in both parties. Then, for each Justice, we extract the percentage of times that they participate in hyperedges of a given majority. This measure indicates the tendency of Justices to vote more often aligned with democrats or republicans, an information summarized in Fig. 3b. We observe Justices that consistently vote with their own party majority (e.g., Justice 3 votes mainly with other democrats, Justice 28 mainly with other republicans), but also cases in which the political party of the Justice is not aligned with the voting behavior expressed by their hyperedges. For example, node 30 (Justice Ruth Bader Ginsburg) is associated with the Democratic Party, but most of her votes align with those of republican Justices. This behavior is captured by Hypergraph-MT, which assigns her a membership more peaked in the community made of republicans and only partially to the one of democrats, as shown in Fig. 3c. Instead, Graph-MT assigns her mostly to the community of democrats. This mismatch between hypergraph information and political affiliation explains the lower value of cosine similarity in Fig. 3a. Similar conclusions can be drawn for node 31 and 15. More generally, the overlapping memberships inferred by Hypergraph-MT match more closely the voting behavior of Justices than those inferred by Graph-MT, as shown in the pie markers in Fig. 3c.

In addition to community structure, Hypergraph-MT outperforms Graph-MT also in the hyperedge prediction task. Figure 4a shows how Hypergraph-MT achieves higher AUC than Graph-MT, in both predicting pairwise and higher-order interactions. This further corroborates the hypothesis that information is lost when decoupling higher-order interactions via clique expansion. This example illustrates why it is critical to consider hypergraphs when hyperedges contain information that can be lost by clique expansion. It also shows the advantage of considering overlapping communities when nodes' behaviors are nuanced and no clear affiliation to one group is expected. As Supreme Court cases span a wide range of topics, we may expect Justices to exhibit a diversity of preferences (and thus voting

behaviors) that cannot be fully captured by a binary political affiliation. Hence, models that consider overlapping communities can provide a variety of patterns that better represents this diversity. Finally, this example also confirms that metadata should be carefully used as “ground-truth” communities, thus encouraging a careful exploration of the relationship between node metadata, information contained in the hyperedges and community structure⁵⁷.

The computational efficiency of Hypergraph-MT. Beyond accuracy, algorithmic efficiency is necessary for a widespread applicability of statistical inference models to large-scale datasets. Hence, we now assess the performance of our model on a variety of systems from different domains, focusing on the analysis of the computational efficiency of Hypergraph-MT as compared to alternative approaches. The higher-order datasets include co-sponsorship and committee memberships data of the U.S. Congress, co-purchasing behavior of customers on Walmart, and clicking activity of users on Trivago (Table 1). Hypergraph-MT and Graph-MT perform similarly in terms of predicting missing hyperedges on most of these datasets, as shown in Fig. 4a. This suggests that in such cases, the information learned from the clique expansion is similar to that contained in a hypergraph representation. While one may be tempted to conclude that using a dyadic method should be favored in these cases, we argue that predictive performance may not be the only metric to use to make this decision. Indeed, time complexity also plays a role here, as many of these datasets have large hyperedges. While we have extensively discussed the efficiency of Hypergraph-MT, one should also consider the cost of running dyadic methods on clique expansion of large data. In fact, this depends on the number of pairs generated in the expansion, a quantity related to both the amount and size of hyperedges. As a result, the size of a graph obtained by clique expansion can become arbitrarily large. For instance, the House bills data results in almost 4×10^5 edges, as opposed to the 4×10^4 hyperedges given by the hypergraph representation. This difference of an order of magnitude has a significant impact in terms of computational complexity. In fact, we observe a difference of an order of magnitude also in the running time of the algorithms, as shown in Fig. 4b, where we plot the time to run the three methods on

each dataset. While for datasets with small hyperedges (e.g., the close-proximity data discussed above) running time is similar for Hypergraph-MT and Graph-MT, we observe significant differences for datasets with larger maximum size D , with Hypergraph-MT being much faster to run. Hypergraph-MT may therefore be the algorithm of choice for large system sizes. See Supplementary Note 2 for further results about the computational complexity of the methods on synthetic data with variable size.

Discussion

Here, we have introduced Hypergraph-MT, a mixed-membership probabilistic generative model for hypergraphs, which proposes a first way to extract the overlapping community organization of nodes in networked systems with higher-order interactions. In addition to detecting communities, our model provides a principled tool to predict missing hyperedges, thus serving as a quantitative evaluation framework for assessing goodness of fit. This feature is particularly useful in the absence of metadata when evaluating community detection schemes. In practice, our model considers an assortative affinity matrix, which makes its algorithmic implementation highly scalable. The computational complexity is also significantly reduced by an efficient routine to compute expensive quantities at low cost in each update, a problem not present in the case of graphs. We have applied our model to a wide variety of social and biological hypergraphs, discussing accuracy in the hyperedge and community structure inference tasks. Moreover, we have showed that Hypergraph-MT outperforms clique expansion methods with respect to running time, making it a suitable solution also for higher-order datasets with large hyperedges.

Our method has a substantial advantage in systems where hyperedges contain important information that can be lost by considering non-higher-order methods on projected dyadic graphs. For instance, it allows quantifying how maximum hyperedge size impacts performance and unveils the presence of critical sizes beyond which higher-order algorithms may significantly outperform dyadic methods, as shown in a Gene-Disease dataset. Hypergraph-MT also has the benefits of being more resilient to the addition of large noisy hyperedges and of being more robust in detecting communities that are more closely aligned with the information carried by hyperedges, as shown in the analysis of the U.S. Justices.

There are natural methodological extensions to further expand the range of applications covered by our model. Here, we have considered an assortative affinity matrix, but alternative formulations could be considered to target different types of structures. The challenge would be to increase flexibility while keeping the dimensionality of the problem under control. Moreover, our model takes in input hyperedges of one type, but there could be multiple types of ways to connect a subset of nodes. Expanding our approach to these cases would be analogous to extend single-layer networks to multi-layer ones. This may be done by suitably defining different types of affinity matrices for each type of high-order interaction, as in ref. 49. Similarly, our model might be extended to extract temporal higher-order communities in the presence of time-varying interactions with memory^{58,59}. Finally, hypergraphs may carry additional information beyond the one contained in hyperedges. This calls for further developments to rigorously incorporate information such as node attributes into the model formulation^{60,61}. While here we have focused on analyzing real-world data, our generative model can also be used to sample synthetic data with hypergraph structure. In particular, our model could prove useful for practitioners interested in utilizing synthetic benchmarks of hypergraphs, allowing a better characterization of higher-order topological properties, including simplicial closure³⁵ and higher-order motifs³⁹. Taken together, Hypergraph-MT provides a fast and scalable tool for inferring the structure of large-scale hypergraphs, contributing to a better

understanding of the networked organization of real-world higher-order systems.

Methods

Inference of Hypergraph-MT

Hypergraph-MT models the likelihood of the hypergraph $\mathbf{A} = \{A_e\}_{e \in \mathcal{E}}$ as:

$$P(\mathbf{A}|\theta) = \prod_{e \in \Omega} e^{-\lambda_e} \frac{\lambda_e^{A_e}}{A_e!}, \tag{7}$$

where $\lambda_e = \sum_k w_{d_e k} \prod_{i \in e} u_{ik}$. The set of latent variables is defined by $\theta = (\mathbf{u}, \mathbf{w})$, where \mathbf{u} is a $N \times K$ -dimensional community membership matrix and \mathbf{w} is a $D \times K$ -dimensional affinity matrix, where $D = \max_{e \in \mathcal{E}} d_e$ is the maximum hyperedge size in the dataset. Each entry w_{dk} represents the density of hyperedges of size d in the community k . Notice, we only consider the assortative regime, to reduce the dimensionality of the affinity tensor \mathbf{w} . The product runs over $\Omega = \{e|e \subseteq \mathcal{V}, d_e \geq 2\}$, that is, the set of all potential hyperedges. In practice, we can reduce this space by considering only the possible hyperedges of a certain size lower or equal than the maximum observed size D . For instance, if the maximum size of interactions in a hypergraph is $D = 4$, then we should not expect to see hyperedges of size 5, and we can define $\Omega = \{e|e \subseteq \mathcal{V}, 2 \leq d_e \leq D\}$.

With this formulation, Hypergraph-MT is a mixed-membership probabilistic generative model for hypergraphs. The main intuition behind it is that a hyperedge is more likely to exist between nodes with the same community membership. In fact, hyperedges in which even a single value $u_{ik} = 0$ appears, are assigned a null probability. The goal is thus to infer the latent variables \mathbf{u} and \mathbf{w} given the observed hypergraph \mathbf{A} .

We infer the parameters using a maximum likelihood approach. Specifically, we maximize the log-likelihood

$$L = - \sum_{e \in \Omega} \sum_k w_{d_e k} \prod_{i \in e} u_{ik} + \sum_{e \in \mathcal{E}} A_e \log \sum_k w_{d_e k} \prod_{i \in e} u_{ik} \tag{8}$$

with respect to $\theta = (\mathbf{u}, \mathbf{w})$, where we neglect the factorial term, which is independent of the parameters. As the summation in the logarithm renders the calculations difficult, we employ a variational approximation using Jensen's inequality, that gives

$$\mathcal{L}(\boldsymbol{\rho}, \theta) = - \sum_{e \in \Omega} \sum_k w_{d_e k} \prod_{i \in e} u_{ik} + \sum_{e \in \mathcal{E}} A_e \sum_k \rho_{ek} \log \left(\frac{w_{d_e k} \prod_{i \in e} u_{ik}}{\rho_{ek}} \right). \tag{9}$$

For each $e \in \mathcal{E}$, we consider a variational distribution ρ_{ek} over the communities k : this is our estimate of the probability that the hyperedge e exists due to the contribution of the community k . The equality holds when

$$\rho_{ek} = \frac{w_{d_e k} \prod_{i \in e} u_{ik}}{\sum_k w_{d_e k} \prod_{i \in e} u_{ik}}. \tag{10}$$

Maximize Eq. (8), is then equivalent to maximize Eq. (9) with respect to both θ and $\boldsymbol{\rho}$. We estimate the parameters by using an expectation-maximization (EM) algorithm, where at each step one updates $\boldsymbol{\rho}$ using Eq. (10) (E-step) and then maximizes $\mathcal{L}(\boldsymbol{\rho}, \theta)$ regarding $\theta = (\mathbf{u}, \mathbf{w})$ by setting partial derivatives to zero (M-step). This procedure is repeated until the log-likelihood converges. The fixed point is a local maximum, but it is not guaranteed to be the global maximum. Therefore, we perform ten runs of the algorithm with different random initialization for θ , taking the fixed point with the largest value of the log-likelihood. For further details, see Supplementary Note 1.

Hyperedge prediction and cross-validation

We assess the performance of our model by measuring the goodness in predicting missing hyperedges. In these experiments, we use a 5-fold cross-validation routine: we divide the dataset into five equal-size groups (folds), selected uniformly at random, and give the models access to four groups (training data) to learn the parameters; this contains 80% of the hyperedges. One then predicts the hyperedges in the held-out group (test set). By varying which group we use as the test set, we get five trials per realization. When we use the baseline Pairs-MT, the training and the test sets are the subsets extracted from the initial ones, containing only the hyperedges with $d_e = 2$. Instead, when we use the baseline Graph-MT, we train the model on the graph obtained from clique expansions of the hyperedges in the training set.

As a performance metric, we measure the area under the receiver-operator characteristic curve (AUC) on the test data, and the final results are averages over the five folds. The AUC is the probability that a random true positive is ranked above a random true negative; thus the AUC is 1 for perfect prediction, and 0.5 for chance. Since the set of all possible hyperedges is large, it is not possible to compute the AUC on the whole training and test sets; hence we proceed with samples. In detail, we fix the number of comparisons we want to evaluate, here 10^3 . We then sample 10^3 values from the non-zero entries (where exist a hyperedge) of the sets, and we save the inferred hyperedge probabilities in a vector R_1 . We sample the same number of values from the zero entries (where do not exist a hyperedge), keeping this set balanced with R_1 in terms of hyperedge size distribution. We save the inferred hyperedge probabilities of this set of entries in a vector R_0 . We then make element-wise comparisons and compute the AUC as

$$\text{AUC} = \frac{\sum(R_1 > R_0) + 0.5 \sum(R_1 = R_0)}{|R_1|}, \quad (11)$$

where $\sum(R_1 > R_0)$ stands for the number of times R_1 has a higher value than R_0 in the element-wise comparisons; and $|R_1| = |R_0|$ is the length of the vector, which is equal to the number of comparisons we fix.

To predict the existence of a hyperedge, we use different approaches according to the structure under analysis. For Hypergraph-MT, the probability of a hyperedge is given by Eq. (1). For Graph-MT, instead, we compute the probability of a hyperedge as the product of the probabilities of each edge of its clique expansion to exist. That is, $P(A_e) = \prod_{(ij) \in e_2} P(A_{ij} > 0)$, where e_2 is the 2-combination set of the hyperedge e . Notice, all the single pairwise interactions have to exist, to have a probability of the hyperedge greater than zero. When evaluating Pairs-MT, we measure the AUC only on the subset of the test set containing edges, i.e., hyperedges with $d_e = 2$. To perform a balanced comparison in this case, we also measure the AUC for both Hypergraph-MT and Graph-MT on this set (pairs), while still training on the whole train set. This provides information on the utility of large hyperedges to predict pairwise interactions.

Description of the datasets

In the main text, we analyze hypergraphs derived from empirical data from various domains, and we provide a summary of study datasets in Table 1. To perform the inference in these datasets, we need to choose the number of communities K . In general, K can be selected using model selection criteria. For instance, one could evaluate the model's predictive performance—for example in the link prediction task—for varying numbers of communities, and then choose the best performing K . Here, for simplicity, we fix the number of communities K equal to the number of classes of a node metadata, aiming to compare the resulting communities with this additional information.

We first analyze four datasets collected by the SocioPatterns collaboration (<http://www.sociopatterns.org>), which describe human

close-proximity contact interactions obtained from wearable sensor data. The High-school dataset describes the interactions between students of nine different classrooms⁶². In the Primary school, nodes are students and teachers and a hyperedge connects groups of people that were all jointly in proximity to one another^{63,64}. Also here, the number of communities reflects the classrooms to which each student belongs, and it includes an additional class for the teachers. The Workplace dataset contains the contacts of individuals of five different departments, measured in an office building in France⁶⁵. Lastly, the Hospital hypergraph collects the interactions between patients, patients and health-care workers (HCWs) and among HCWs in a hospital ward in France⁶⁶. The number of communities corresponds then to the number of roles in the ward.

We then analyze the Gene-Disease dataset, that describes the gene-disease associations provided by expert curated resources (e.g., UNIPROT, CTI)⁶⁷. Nodes correspond to genes, and each hyperedge is the set of genes associated with a disease. We keep only the genes with a non-nan value of the Disease Pleiotropy Index (DPI), a quantity that considers if the diseases associated with the gene are similar among them and belong to the same disease class or belong to different disease classes. We use this attribute to fix the number of communities because it may indicate the different behaviors of the genes in the datasets. Moreover, we keep hyperedges with size $2 \leq d_e \leq 25$.

The second case study in the main text presents the analysis of the Justice hypergraph constructed from the data in <http://scdb.wustl.edu/about.php>. This dataset records all the votes expressed by the justices of the Supreme Court in the U.S. from 1946 to 2019 case by case. Nodes correspond to justices, and each hyperedge is the set of justices that expressed the same vote in a case. The number of communities corresponds to the number of political parties, i.e., Democrat and Republican.

The following datasets have been downloaded from <https://www.cs.cornell.edu/~arb/data/>. We analyze hypergraphs created from U.S. congressional bill co-sponsorship data, where nodes correspond to congresspersons and hyperedges correspond to the sponsor and all cosponsors of a bill in either the House of Representatives (House bills) or the Senate (Senate bills)^{45,68,69}. We also use two datasets from the U.S. Congress in the form of committee memberships^{45,70}. Each hyperedge is a committee in a meeting of Congress, and each node again corresponds to a member of the House (House committees) or a senator (Senate committees). A node is contained in a hyperedge if the corresponding legislator was a member of the committee during the specified meeting of Congress. In all these congressional datasets, the node labels give the political parties of the members, thus all of them have $K = 2$. For these datasets, we run the model with different values of $D = 2, \dots, 25$ and choose the best value among them.

In addition to the congressional datasets, we analyze the Walmart hypergraph⁷¹. Here, each node is a product, and a hyperedge connects a set of products that were co-purchased by a customer in a single shopping trip. We fix the number of communities equal to the product category labels. Lastly, we analyze the Trivago dataset⁴⁵. Nodes correspond to hotels listed at trivago.com, and each hyperedge corresponds to a set of hotels whose website was clicked on by a user of Trivago within a browsing session. For each hotel, the node label gives the country in which it is located, and we fix K based on this information. For Walmart and Trivago, we consider a subset of the hypergraph to reduce the sparsity, as done in ref. 45. The c -core of a hypergraph \mathcal{H} is defined as the largest subhypergraph \mathcal{H}_c such that all nodes in \mathcal{H}_c have size at least c . For Walmart, we use the 3-core hypergraph, and for Trivago, we work with the 5-core hypergraph.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets used in the paper are publicly available from their sources listed in the Methods section.

Code availability

An open-source algorithmic implementation of the model is publicly available and can be found at <https://github.com/mcontisc/Hypergraph-MT>.

References

- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
- Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320 (2019).
- Klamt, S., Haus, U.-U. & Theis, F. Hypergraphs and cellular networks. *PLOS Computat. Biol.* **5**, e1000385 (2009).
- Petri, G. et al. Homological scaffolds of brain functional networks. *J. R. Soc. Interface* **11**, 20140873 (2014).
- Giusti, C., Ghrist, R. & Bassett, D. S. Two's company, three (or more) is a simplex. *J. Comput. Neurosci.* **41**, 1–14 (2016).
- Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
- Grilli, J., Barabás, G., Michalska-Smith, M. J. & Allesina, S. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017).
- Gao, Y. et al. Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans. Image Process.* **22**, 363–376 (2012).
- Cencetti, G., Battiston, F., Lepri, B. & Karsai, M. Temporal properties of higher-order interactions in social networks. *Sci. Rep.* **11**, 1–10 (2021).
- Patania, A., Petri, G. & Vaccarino, F. The shape of collaborations. *EPJ Data Sci.* **6**, 1–16 (2017).
- Berge, C. *Graphs and Hypergraphs* (North-Holland Pub. Co., 1973).
- Battiston, F. et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
- Torres, L., Blevins, A. S., Bassett, D. & Eliassi-Rad, T. The why, how, and when of representations for complex systems. *SIAM Rev.* **63**, 435–485 (2021).
- Battiston, F. & Petri, G. *Higher-Order Systems* (Springer, 2022).
- Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
- Schaub, M. T., Benson, A. R., Horn, P., Lippner, G. & Jadbabaie, A. Random walks on simplicial complexes and the normalized hodge 1-laplacian. *SIAM Rev.* **62**, 353–391 (2020).
- Carletti, T., Battiston, F., Cencetti, G. & Fanelli, D. Random walks on hypergraphs. *Phys. Rev. E* **101**, 022308 (2020).
- Bick, C., Ashwin, P. & Rodrigues, A. Chaos in generically coupled phase oscillator networks with nonpairwise interactions. *Chaos* **26**, 094814 (2016).
- Skardal, P. S. & Arenas, A. Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.* **3**, 1–6 (2020).
- Millán, A. P., Torres, J. J. & Bianconi, G. Explosive higher-order kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.* **124**, 218301 (2020).
- Lucas, M., Cencetti, G. & Battiston, F. Multiorder laplacian for synchronization in higher-order networks. *Phys. Rev. Res.* **2**, 033410 (2020).
- Gambuzza, L. V. et al. Stability of synchronization in simplicial complexes. *Nat. Commun.* **12**, 1–13 (2021).
- Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nat. Commun.* **10**, 1–9 (2019).
- Chowdhary, S., Kumar, A., Cencetti, G., Iacopini, I. & Battiston, F. Simplicial contagion in temporal higher-order networks. *J. Phys.* **2**, 035019 (2021).
- Neuhäuser, L., Mellor, A. & Lambiotte, R. Multibody interactions and nonlinear consensus dynamics on networked systems. *Phys. Rev. E* **101**, 032310 (2020).
- Alvarez-Rodriguez, U. et al. Evolutionary dynamics of higher-order interactions in social networks. *Nat. Human Behav.* **5**, 586–595 (2021).
- Kovalenko, K. et al. Growing scale-free simplices. *Commun. Phys.* **4**, 1–9 (2021).
- Millán, A. P., Ghorbanchian, R., Defenu, N., Battiston, F. & Bianconi, G. Local topological moves determine global diffusion properties of hyperbolic higher-order networks. *Phys. Rev. E* **104**, 054302 (2021).
- Courtney, O. T. & Bianconi, G. Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Phys. Rev. E* **93**, 062311 (2016).
- Young, J.-G., Petri, G., Vaccarino, F. & Patania, A. Construction of and efficient sampling from the simplicial configuration model. *Phys. Rev. E* **96**, 032312 (2017).
- Chodrow, P. S. Configuration models of random hypergraphs. *J. Complex Netw.* **8**, cnaa018 (2020).
- Patania, A., Vaccarino, F. & Petri, G. Topological analysis of data. *EPJ Data Sci.* **6**, 1–6 (2017).
- Sizemore, A. E., Phillips-Cremins, J. E., Ghrist, R. & Bassett, D. S. The importance of the whole: topological data analysis for the network neuroscientist. *Netw. Neurosci.* **3**, 656–673 (2019).
- Young, J.-G., Petri, G. & Peixoto, T. P. Hypergraph reconstruction from network data. *Commun. Phys.* **4**, 1–11 (2021).
- Benson, A. R., Abebe, R., Schaub, M. T., Jadbabaie, A. & Kleinberg, J. Simplicial closure and higher-order link prediction. *Proc. Natl Acad. Sci. USA* **115**, E11221–E11230 (2018).
- Krishnagopal, S. & Bianconi, G. Spectral detection of simplicial communities via hodge laplacians. *Phys. Rev. E* **104**, 064303 (2021).
- Benson, A. R. Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.* **1**, 293–312 (2019).
- Tudisco, F. & Higham, D. J. Node and edge nonlinear eigenvector centrality for hypergraphs. *Commun. Phys.* **4**, 1–10 (2021).
- Lotito, Q. F., Musciotto, F., Montresor, A. & Battiston, F. Higher-order motif analysis in hypergraphs. *Commun. Phys.* **5**, 79 (2022).
- Musciotto, F., Battiston, F. & Mantegna, R. N. Detecting informative higher-order interactions in statistically validated hypergraphs. *Commun. Phys.* **4**, 1–9 (2021).
- Wolf, M. M., Klinvex, A. M. & Dunlavy, D. M. 2016 *IEEE High Performance Extreme Computing Conference (HPEC)* 1–7 (IEEE, 2016).
- Vazquez, A. Finding hypergraph communities: a bayesian approach and variational solution. *J. Stat. Mech.* **2009**, P07006 (2009).
- Carletti, T., Fanelli, D. & Lambiotte, R. Random walks and community detection in hypergraphs. *J. Phys.* **2**, 015011 (2021).
- Eriksson, A., Edler, D., Rojas, A., de Domenico, M. & Rosvall, M. How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun. Phys.* **4**, 1–12 (2021).
- Chodrow, P. S., Veldt, N. & Benson, A. R. Generative hypergraph clustering: From blockmodels to modularity. *Sci. Adv.* **7**, eabh1303 (2021).
- Chodrow, P., Eikmeier, N. & Haddock, J. Nonbacktracking spectral clustering of nonuniform hypergraphs. Preprint at <https://arxiv.org/abs/2204.13586> (2022).
- Zhou, D., Huang, J. & Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **19**, 1601–1608 (2006).
- Ball, B., Karrer, B. & Newman, M. E. Efficient and principled method for detecting communities in networks. *Phys. Rev. E* **84**, 036103 (2011).

49. De Bacco, C., Power, E. A., Larremore, D. B. & Moore, C. Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
50. Goldenberg, A., Zheng, A. X., Fienberg, S. E. & Airoldi, E. M. A survey of statistical network models. *Found. Trends Mach. Learn.* **2**, 129–233 (2010).
51. Fortunato, S. & Hric, D. Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44 (2016).
52. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivelä, M. Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eaax7310 (2020).
53. Safdari, H., Contisciani, M. & De Bacco, C. Generative model for reciprocity and community detection in networks. *Phys. Rev. Res.* **3**, 023209 (2021).
54. Contisciani, M., Safdari, H. & De Bacco, C. Community detection and reciprocity in networks by jointly modelling pairs of edges. *J. Complex Netw.* **10**, cnac034 (2022).
55. Safdari, H., Contisciani, M. & De Bacco, C. Reciprocity, community detection, and link prediction in dynamic networks. *J. Phys.* **3**, 015010 (2022).
56. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser B (Methodological)* **39**, 1–22 (1977).
57. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
58. Scholtes, I. et al. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nat. Commun.* **5**, 1–9 (2014).
59. Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D. & Lambiotte, R. Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Commun.* **5**, 1–13 (2014).
60. Contisciani, M., Power, E. A. & De Bacco, C. Community detection with node attributes in multilayer networks. *Sci. Rep.* **10**, 1–16 (2020).
61. Newman, M. E. & Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **7**, 1–11 (2016).
62. Mastrandrea, R., Fournet, J. & Barrat, A. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* **10**, e0136497 (2015).
63. Gemmetto, V., Barrat, A. & Cattuto, C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infect. Dis.* **14**, 1–10 (2014).
64. Stehlé, J. et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE* **6**, e23176 (2011).
65. Génois, M. et al. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Netw. Sci.* **3**, 326–347 (2015).
66. Vanhems, P. et al. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE* **8**, e73970 (2013).
67. Piñero, J. et al. The disgenet knowledge platform for disease genomics: 2019 update. *Nucl. Acids Res.* **48**, D845–D855 (2020).
68. Fowler, J. H. Connecting the congress: A study of cosponsorship networks. *Polit. Anal.* **14**, 456–487 (2006).
69. Fowler, J. H. Legislative cosponsorship networks in the us house and senate. *Soc. Netw.* **28**, 454–465 (2006).
70. Stewart, C. III & Woon, J. Congressional Committee assignments, 103rd to 114th Congresses, 1993–2017: House, *Technical Report*, MIT mimeo (2008).
71. Amburg, I., Veldt, N. & Benson, A. *Clustering in Graphs and Hypergraphs with Categorical Edge Labels*. 706–717 (Association for Computing Machinery, 2020).

Acknowledgements

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting M.C.; C.D.B. and M.C. were supported by the Cyber Valley Research Fund. F.B. acknowledges support from the Air Force Office of Scientific Research under award number FA8655-22-1-7025. The authors thank Philip S. Chodrow and Nate Veldt for useful discussions.

Author contributions

M.C. and C.D.B. developed the algorithm and performed the experiments. M.C., F.B., and C.D.B. all conceived the research, analyzed the results and wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-34714-7>.

Correspondence and requests for materials should be addressed to Martina Contisciani or Caterina De Bacco.

Peer review information *Nature Communications* thanks Filippo Radicchi, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022



NETWORK SCIENCE

Community detection in large hypergraphs

Nicolò Ruggeri^{1,2*}, Martina Contisciani¹, Federico Battiston³, Caterina De Bacco^{1*}

Hypergraphs, describing networks where interactions take place among any number of units, are a natural tool to model many real-world social and biological systems. Here, we propose a principled framework to model the organization of higher-order data. Our approach recovers community structure with accuracy exceeding that of currently available state-of-the-art algorithms, as tested in synthetic benchmarks with both hard and overlapping ground-truth partitions. Our model is flexible and allows capturing both assortative and disassortative community structures. Moreover, our method scales orders of magnitude faster than competing algorithms, making it suitable for the analysis of very large hypergraphs, containing millions of nodes and interactions among thousands of nodes. Our work constitutes a practical and general tool for hypergraph analysis, broadening our understanding of the organization of real-world higher-order systems.

INTRODUCTION

Over the last decades, most relational data, from biological to social systems, have found a successful representation in terms of networks, where nodes describe the basic units of the system, and link their pairwise interactions (1). Nevertheless, such a modeling approach cannot properly encode the presence of group interactions, describing associations among three or more system units at a time (2–5). Such higher-order interactions have been observed in a wide variety of systems, including collaboration networks (6), cellular networks (7), drug recombination (8), human (9) and animal (10) face-to-face interactions, and structural and functional mapping of the human brain (11–13). In addition, the higher-order organization of many interacting systems is associated with the generation of new phenomena and collective behavior across many different dynamical processes, such as diffusion (14), synchronization (15–20), spreading (21–23), and evolutionary games (24–26).

Networked systems with higher-order interactions are better described by different mathematical frameworks from networks, such as hypergraphs, where hyperedges encode interactions among an arbitrary number of system units (2, 27). In the last few years, several tools have been developed for higher-order network analysis. These include higher-order centrality scores (28, 29), clustering (30), and motif analysis (31, 32), as well as higher-order approaches to network backboning (33, 34), link prediction (35), and methods to reconstruct nondyadic relationships from pairwise interaction records (36). A variety of approaches have been suggested to detect communities in hypergraphs, including nonparametric methods with hypergraphons (37), tensor decompositions (38), latent space distance models (39), latent class models (40), flow-based algorithms (41, 42), spectral clustering (43–45), and spectral embeddings (46). A different line of works focuses on deriving theoretical detectability limits (47–49).

Recently, statistical inference frameworks have been proposed to capture in a principled way the mesoscale organization of hypergraphs (35, 50, 51). Despite their success, current approaches

suffer from a number of notable drawbacks. For instance, the method in (51) is restricted to using very small hypergraphs and hyperedges, due to its high computational complexity. Also, the approach in (50) suffers from a high computational complexity in the general case and needs to make strong assumptions to scale to real-life datasets. Finally, the model in (35) is constrained to work only with assortative community structures.

Here, we propose a framework to model the organization of higher-order systems. Our method allows detecting communities in hypergraphs with accuracy exceeding that of state-of-the-art approaches, in the cases of both hard and mixed community assignments, as we show on synthetic benchmarks with known ground-truth partitions. Furthermore, its flexibility allows capturing general configurations that could not be previously studied, such as disassortative community interactions.

Finally, overcoming the computational thresholds of previous methods, our model is extremely efficient, making it suitable to study hypergraphs containing millions of nodes and interactions among thousands of system units not accessible to alternative tools. We illustrate the advantages of our approach through a variety of experiments on synthetic and real data. Our results showcase the wide applicability of the proposed method, contributing to broaden our understanding of the organization of higher-order real-world systems.

GENERATIVE MODEL

A hypergraph consists of a set of nodes $V = \{1, \dots, N\}$ and a set of hyperedges E . Each hyperedge e is a subset of V , representing a higher-order interaction between a number $|e|$ of nodes. We denote by D the maximum possible hyperedge size, which can be arbitrarily imposed up to a maximum value of $D = N$, and Ω the set of all possible hyperedges among nodes in V . We represent the hypergraph via an adjacency vector $\mathbf{A} \in \mathbb{N}^\Omega$, with entry A_e being the weight of $e \in \Omega$. We assume the weights A_e to be nonnegative and discrete. For real-world systems, \mathbf{A} is typically sparse. The number $|E|$ of nonzero entries is typically linear in N , and thus much smaller than the dimension $|\Omega|$.

We model hypergraphs probabilistically, assuming an underlying arbitrary community structure with K overlapping groups, similarly to a mixed-membership stochastic block model. Each node i

¹Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany. ²Department of Computer Science, ETH, 8004 Zürich, Switzerland. ³Department of Network and Data Science, Central European University, 1100 Vienna, Austria.

*Corresponding author. Email: nicolo.ruggeri@tuebingen.mpg.de (N.R.); caterina.debacco@tuebingen.mpg.de (C.D.B.)

can potentially belong to multiple groups, as specified by a K -dimensional membership vector \mathbf{u}_i with nonnegative entries. We collect all the membership assignments in a $N \times K$ matrix u . The density of interactions within and between communities is regulated by a symmetric nonnegative $K \times K$ affinity matrix w . These two main parameters, u and w , control the Poisson distributions of the hyperedge weights

$$p(A_e; u, w) = \text{Pois}\left(A_e; \frac{\lambda_e}{\kappa_e}\right) \tag{1}$$

where

$$\begin{aligned} \lambda_e &= \sum_{i < j: i, j \in e} \mathbf{u}_i^T w u_j \\ &= \sum_{i < j: i, j \in e} \sum_{k, q=1}^K u_{ik} u_{jq} w_{kq} \end{aligned} \tag{2}$$

Here, $\kappa_e = \kappa_{|e|}$ is a normalization factor that solely depends on the hyperedge size $|e|$. We develop our theory for a general form of κ_n . While in principle any choice $\kappa_n > 0$ is possible, in our experiments we use the form $\kappa_n = \frac{n(n-1)}{2} \binom{N-2}{n-2}$, for every hyperedge of size n (52). Because of the fact that $\kappa_2 = 1$, if the hypergraph contains only pairwise interactions our model is similar to existing mixed-membership block models for dyadic networks (53, 54). Intuitively, given two nodes i, j , the term $\binom{N-2}{n-2}$ normalizes for the number of possible choices of the remaining $n-2$ nodes in the hyperedge. The term $n(n-1)/2$ averages among the number of possible pairwise interactions among the n nodes in the hyperedge. Note that previous generative models for hypergraphs were limited to detect only assortative community interactions (35, 50). By contrast, in our model, each entry w_{kq} distinctly specifies the strength of the interactions between each k, q community pair. Hence, for the first time, our method allows encoding more general community structures, without the need to impose a priori assumptions to ensure computational and theoretical feasibility. In particular, the bilinear form in Eq. 2 allows for a tractable and scalable inference, regardless of the structure of w . Another relevant feature of the model is that the size of the affinity matrix w does not vary with maximum hyperedge size D nor with the number of hyperedges, making it memory efficient also for hypergraphs with large interactions. We name our model Hy-MMSBM, for hypergraph mixed-membership stochastic block model, and provide an open-source implementation at <http://github.com/nickruggeri/Hy-MMSBMgithub.com/nickruggeri/Hy-MMSBM>. We have also incorporated our algorithm inside the open-source library Hypergraphx (55).

INFERENCE

Optimization procedure

In real-life scenarios, practitioners observe a list of hyperedges, encoded in the vector \mathbf{A} , and aim to learn the node memberships u and affinity matrix w that best fit the data. To this end, we start by considering the likelihood of \mathbf{A} given the parameters $\theta = (u, w)$. Using Eqs. 1 and 2, this is given by

$$p(\mathbf{A} | \theta) = \prod_{e \in \Omega} \text{Pois}\left(A_e; \frac{\lambda_e}{\kappa_e}\right) \tag{3}$$

where the hyperedge weights are assumed to be conditionally independent given (u, w) . Its logarithm is given by

$$\begin{aligned} \log p(\mathbf{A} | \theta) &= \sum_{e \in \Omega} -\frac{1}{\kappa_e} \sum_{i < j \in e} \mathbf{u}_i^T w u_j \\ &\quad + \sum_{e \in E} A_e \log \sum_{i < j \in e} \mathbf{u}_i^T w u_j \end{aligned} \tag{4}$$

where we discarded constant terms not depending on the parameters. The first summation over $|\Omega|$ terms appears intractable due to the exploding size of the configuration space. However, one important feature of our model is that this high dimensionality can be treated analytically, as the likelihood conveniently simplifies. The summand $\sum_{e \in \Omega} -\frac{1}{\kappa_e} \sum_{i < j \in e} \mathbf{u}_i^T w u_j$ is simply taking the interaction term $\mathbf{u}_i^T w u_j$ as many times as it appears in all the possible hyperedges, each weighted by the factor $1/\kappa_e$. This reasoning yields the count $C = \sum_{n=2}^D \frac{1}{\kappa_n} \binom{N-2}{n-2}$ and the following simplified log-likelihood

$$\begin{aligned} \log p(\mathbf{A} | \theta) &= -C \sum_{i < j \in V} \mathbf{u}_i^T w u_j \\ &\quad + \sum_{e \in E} A_e \log \sum_{i < j \in e} \mathbf{u}_i^T w u_j \end{aligned} \tag{5}$$

obtaining a tractable sum of terms. To maximize Eq. 5 with respect to u and w , we use a standard variational approach via Jensen’s inequality $\log \mathbb{E}[x] \geq \mathbb{E}[\log x]$ to lower bound the second summand as

$$\begin{aligned} \sum_{e \in E} A_e \log \sum_{i < j \in e} \mathbf{u}_i^T w u_j &\geq \\ \sum_{e \in E} A_e \sum_{i < j \in e} \sum_{k, q=1}^K \rho_{ijkq}^{(e)} \log \left(\frac{u_{ik} u_{jq} w_{kq}}{\rho_{ijkq}^{(e)}} \right) \end{aligned} \tag{6}$$

Here, the variational distribution is specified by the $\rho_{ijkq}^{(e)}$ values, which can be any configuration of strictly positive probabilities such that $\sum_{i < j \in e} \sum_{k, q=1}^K \rho_{ijkq}^{(e)} = 1$. The equality in Eq. 6 is achieved when

$$\rho_{ijkq}^{(e)} = \frac{u_{ik} u_{jq} w_{kq}}{\sum_{i < j \in e} \sum_{k, q=1}^K u_{ik} u_{jq} w_{kq}} = \frac{u_{ik} u_{jq} w_{kq}}{\lambda_e} \tag{7}$$

Hence, maximizing $\log p(\mathbf{A} | \theta)$ is equivalent to maximizing

$$\begin{aligned} \mathcal{L}(u, w, \rho) &= -C \sum_{i < j \in V} \mathbf{u}_i^T w u_j \\ &\quad + \sum_{e \in E} A_e \sum_{i < j \in e} \sum_{k, q=1}^K \rho_{ijkq}^{(e)} \log \left(\frac{u_{ik} u_{jq} w_{kq}}{\rho_{ijkq}^{(e)}} \right) \end{aligned}$$

with respect to both (u, w) and ρ . This can be done by alternating between updating ρ and (u, w) , as in the expectation-maximization (EM) algorithm.

The update for $\theta \in \{u, w\}$ is obtained by setting the partial derivative $\partial \mathcal{L}(\theta, \rho) / \partial \theta$ to 0, which yields the following expressions

$$u_{ik} = \frac{\sum_{e \in E: i \in e} A_e \rho_{ik}^{(e)}}{C \sum_q w_{kq} \sum_{j \neq i \in V} u_{jq}} \tag{8}$$

$$w_{kq} = \frac{\sum_{e \in E} A_e \rho_{kq}^{(e)}}{C \sum_{i < j \in V} u_{ik} u_{jq}} \tag{9}$$

The terms $\rho_{ik}^{(e)}, \rho_{kq}^{(e)}$ are defined as

$$\rho_{ik}^{(e)} = \sum_{j \in e; j \neq i} \sum_q \rho_{ijkq}^{(e)}$$

$$\rho_{kq}^{(e)} = \sum_{i < j \in e} \rho_{ijkq}^{(e)}$$

and obtained after updating $\rho_{ijkq}^{(e)}$ according to Eq. 7. These updates presented in this section are based on maximum likelihood estimation, where we do not set any prior for (u, w) . However, we can get maximum a posteriori estimates (MAP) with similar derivations and complexity by arbitrarily setting prior distributions for the parameters, as we show in the Supplementary Materials [Appendix Maximum-a-Posteriori (MAP) estimation]. We comment on how to obtain efficient matrix operations that implement the updates in Eqs. 8 and 9 in the “Practical implementation and efficiency” section.

Identifiability, interpretation, and theoretical implications

In the following, we make some observations on relevant aspects regarding the identifiability, interpretation, and theoretical implications of the proposed generative model. First, the log-likelihood in Eq. 5 is invariant under permutations of the groups and under the rescaling $u \rightarrow c u$ and $w \rightarrow w/c^2$, for any constant $c > 0$. This observation may raise questions about identifiability of the parameters. However, both permutation and rescaling do not change the composition of the communities or the relative magnitude of the entries of w ; thus, the mesoscale structure is not affected by them. Nevertheless, one can easily make the model identifiable by setting a prior probability on w and considering MAP estimates (see Appendix Identifiability in the Supplementary Materials for details).

Second, for similar invariance reasons, the constant C can be neglected and absorbed after convergence, by either rescaling $u' = \sqrt{C} u$ or $w' = C w$. While the forms of the rescaling constants κ_e play no role during inference, as they only enter the updates through the C term, they do instead affect the generative process when sampling hypergraphs from it (52). For instance, calculations similar to those in the Supplementary Materials (Appendix Average degree) allow getting a closed-form expression for the average weighted degree when only considering interactions of size k . The resulting formula $\mathbb{E}[d_k^w] = \binom{N-2}{k-2} \frac{k}{\kappa_k N} \sum_{i < j \in V} \mathbf{u}_i^T w \mathbf{u}_j$ shows that rescaling the constant κ_k translates into a rescaling of the average degree. Similar considerations apply to the expected number of hyperedges of a given size and show that the normalization constants κ_e play an important role in determining the expected statistics of the model and hence of the samples they produce. Generally, the sampling procedure from the generative model in Eq. 3, allows determining the degree sequence (i.e., the degree array of the single nodes) as well as the size sequence (i.e., the count of hyperedges for every specified size), which depend on the Poisson parameters and hence on the κ_e normalizers. Alternatively, the sampling procedure from our generative model can be conditioned to respect such sequences (52).

Third, it is possible to obtain the analytical expressions of the expected degree of a node i , which evaluates to

$$\mathbb{E}[d_i^w] = \sum_{e \in \Omega; i \in e} \mathbb{E}[A_e]$$

$$= C \mathbf{u}_i^T w \sum_{j \in V; j \neq i} \mathbf{u}_j + C' \sum_{j < m \in V; j, m \neq i} \mathbf{u}_j^T w \mathbf{u}_m$$

where $C' = \sum_{d=3}^D \frac{\binom{N-3}{d-3}}{\kappa_d}$ is a constant similar to C (see Appendix Average degree in the Supplementary Materials). This expression has a relevant interpretation, as it reveals a fundamental difference between simple networks and higher-order systems. Since in dyadic systems $C' = 0$, we can think of the rightmost summand as a term contributing only to higher-order interactions, while the leftmost one is a shift of the expected degree coming from binary interactions only. One can also observe an analogy with networks of interactions in physical systems. In this context, the leftmost summand can be seen as a mean-field acting on node i in a cavity system where the node is hypothetically removed, while the rightmost term acts as a background field generated by all interactions involving any pair of nodes that does not include node i . This background term is peculiar to higher-order systems, as remarked above. Its presence has a relevant effect of building higher-order interactions between nodes in different groups. This can be illustrated with a simple example of a system with assortative w and node i belonging to a different community than all the other nodes. While the leftmost summand yields expected degree zero in dyadic systems, the background field allows i to form on average nonzero edges. Intuitively, this difference is due to the bilinear form in Eq. 2, which allows observing hyperedges that are not completely homogeneous, where there could be a minor fraction of nodes that are in different communities than the majority. Notice that such a generation, allowing for mixed hyperedges, is a desirable feature. On the one hand, it is appropriate to model contexts where individuals have multiple preferences and thus are expected to belong to multiple groups. On the other hand, recent work (56) proves the combinatorial unfeasibility of hypergraphs where all nodes exhibit majority homophily—implying rather uniform hyperedges contained in single communities—and encourages the development of more flexible generative models.

Practical implementation and efficiency

From an optimization perspective, the EM algorithm starts by initializing u and w at random and then repeatedly alternating between the Eq. 8 and Eq. 9 updates until convergence of $\mathcal{L}(u, w, \rho)$. This does not guarantee to reach the global optimum, but only a local one. In practice, one runs the algorithm several times, each time from a different random initialization, and outputs the parameters corresponding to the realization with highest log-likelihood $\mathcal{L}(u, w, \rho)$. We provide a pseudocode description of the whole inference procedure in Algorithm 1. For all our experiments, we perform MAP inference on the affinity w , setting a factorized exponential prior with rate 1, and maximum likelihood inference on the assignment u . This choice corresponds to the half-Bayesian model presented in the Supplementary Materials [Appendices Maximum-1-Posteriori (MAP) estimation and Identifiability]. The updates have linear computational cost, obtained by exploiting the sparsity of most real-world datasets with efficient matrix operations, as we show in Appendix Computational considerations in the

Downloaded from https://www.science.org on July 12, 2023

Supplementary Materials. Overall, the complexity scales as $O(NK + |E|)$, allowing to tackle inference on hypergraphs whose number of nodes and hyperedges was previously prohibitive (see the “Modeling of real data” section). Another advantage of our inference procedure is that it is stable and reliable for extremely large hyperedges. Because of computational and numerical constraints, previous models were also limited to considering hyperedges with maximal size $D = 25$ (35, 50). As we illustrate in the “Modeling of real data” section with an Amazon and a Gene-Disease dataset, large interactions (respectively $D = 9350$ and $D = 1074$) should not be neglected as they provide useful information and substantially boost the quality of inference.

Algorithm 1: Hy-MMSBM EM inference

Input: Hypergraph A , training rounds r
Result: Inferred parameters (u, w)

```

1 BestLoglik =  $-\infty$ 
2 BestParams = None
  > Train model  $r$  times and choose
  > realization with best likelihood
3 for  $t = 1, \dots, r$  do
  > Initialize at random
4   $u, w \leftarrow \text{init}(u, w)$ 
  > convergence is attained for a max
  number of EM steps, or below a
  certain change in parameter values
5  while not converged do
6     $u \leftarrow \text{update}(u)$  Eq. (8)
7     $w \leftarrow \text{update}(w)$  Eq. (9)
8  end
9   $L = \text{loglik}(u, w)$  Eq. (5)
10 if  $L > \text{BestLoglik}$  then
11   BestLoglik  $\leftarrow L$ 
12   BestParams  $\leftarrow (u, w)$ 
13 end
14 end

```

RECOVERY OF GROUND-TRUTH COMMUNITIES

A standard way to assess the effectiveness of a community detection algorithm is to check if the inferred node memberships match those of a given ground truth. Such ground truth is generally not available for real-world systems (57), while it can be imposed as a planted configuration for synthetic data. For this reason, we consider a recently developed sampling method to produce structured synthetic hypergraphs with flexible structures specified in input (52). For further details, see Appendix Recovery of community assignments in the Supplementary Materials.

In Fig. 1, we generate hypergraphs with an underlying diagonal affinity matrix w (assortative structure) and show the recovery performance for the cases of hard (left) and mixed-membership (right) community assignments. The detailed description of the data generation process is provided in Appendix Recovery of community assignments in the Supplementary Materials. We compare our approach with Hypergraph-MT (35), an inference algorithm designed to detect overlapping community assignments and assortative interactions; Spectral Clustering (43), which recovers hard

communities via hypergraph cut optimization; and Hypergraph AON-MLL (50), which performs a modularity-like optimization based on a Poisson generative model with hard memberships. For our comparisons, we compute the cosine similarity between the ground truth and the inferred communities, which is appropriate to measure the similarity for both hard and mixed-membership vectors. A value of zero represents no similarity, while a value of one is attained by completely overlapping vectors. In both cases, we find that our model successfully recovers the ground-truth communities as more information is made available in terms of hyperedges of increasing sizes. This is somehow expected because the generating process of these data reflects the one of our method, and is a sanity check of our maximum likelihood approach. Spectral Clustering and Hypergraph-MT attain comparable cosine similarity scores on hard-membership data (left), while their performances differ when detecting mixed memberships (right), with Hypergraph-MT performing better. This is because Spectral Clustering performs an approximate combinatorial search and can only recover hard communities, while Hypergraph-MT allows for overlapping communities via maximum likelihood inference. The low performance of Hypergraph AON-MLL is explained by its generative assumptions. AON-MLL assigns the same probability to all the hyperedges containing nodes from more than one community. As most of the hyperedges in this synthetic data are made of nodes from more than one community, the recovery of hypergraph modularity on such systems is close to random. Altogether, such results highlight the effectiveness of the inference procedure, making our model suitable for networked systems with higher-order interactions. Although relevant, the results in Fig. 1 are just one possible comparison among algorithms with different generative assumptions. Such assumptions are expected to yield better or worse results depending on the data, and in general, the no-free-lunch theorem implies that no algorithm will consistently outperform all others on all types of data. As a case for this argument, in Appendix Additional experiments on ground truth recovery in the Supplementary Materials, we present additional results on different synthetic data.

DETECTABILITY OF COMMUNITY CONFIGURATION

Previous inference algorithms rely on the strong assumption of assortative community interactions, hampering their ability to model more complex mesoscale patterns observed in the real world. By contrast, our model allows detecting a variety of different regimes, as it assumes a more flexible w .

Here, we investigate the detection—and detectability—of different assortative and disassortative community structures in hypergraphs, generalizing previous work on pairwise systems (58). In particular, we generate hypergraphs with hard community assignments and different community interactions. We take affinity matrices w with diagonal values c_{in} and out-diagonal values c_{out} , and vary both c_{in} and the ratio $c_{\text{out}}/c_{\text{in}}$. By fixing the value of $c_{\text{out}}/c_{\text{in}}$, we expect higher detectability with increasing c_{in} , as this term regulates the expected degree and consequently the information contained in the data. On the contrary, for a fixed value of c_{in} , we expect the disassortative model to attain better recovery as the ratio $c_{\text{out}}/c_{\text{in}}$ increases, due to the stronger intercommunity interactions. Details on data generation are provided in Appendix Detection of community structure in the Supplementary Materials.

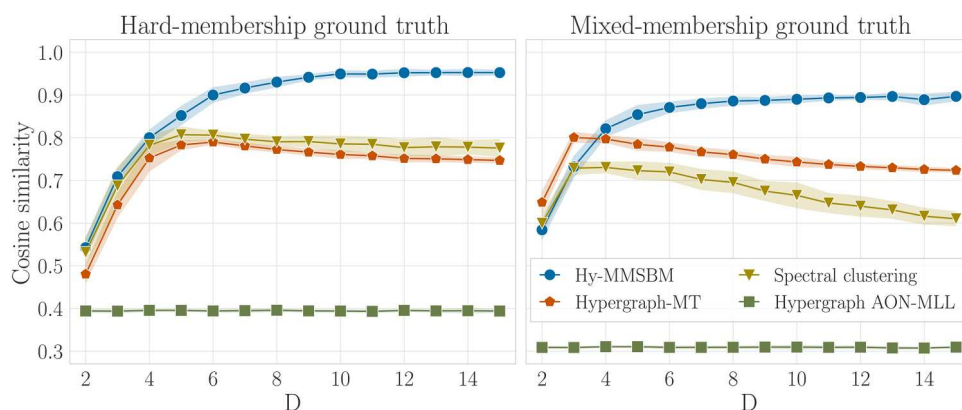


Fig. 1. Recovery of ground-truth community assignments. We measure the cosine similarity between the ground truth and the inferred assignments. We vary the maximum hyperedge size D in synthetic data and study the cases of hard (left) and mixed (right) ground-truth memberships. When information is scarce, represented by few hyperedges of small maximum size D , our method is comparable to the most efficient approaches currently available. However, as larger hyperedges are considered, our method outperforms competing algorithms, on both hard and mixed-membership planted partitions.

We compare the log-likelihoods obtained by the model when the affinity matrix w is initialized as diagonal or full, which we refer to as assortative and disassortative, respectively. Notice that the multiplicative updates in Eq. 9 guarantee that, if w is initialized as diagonal, it will remain as such during training. It is also possible that a full matrix will converge to diagonal during inference. Nonetheless, the strong bias of a diagonal initialization restricts the parameter space of the assortative model, facilitating the convergence to better optima for the detection of assortative structures.

Given the log-likelihood of the assortative (\mathcal{L}_a) and disassortative (\mathcal{L}_d) models, we measure the difference $\mathcal{L}_a - \mathcal{L}_d$ while varying the values of c_{in} and c_{out}/c_{in} . Positive values denote stronger performance of the assortative model, as its likelihood is higher, while negative values favor the disassortative one. We observe that the assortative model attains higher likelihood for low values of c_{out}/c_{in} , when within-community interactions are stronger, as shown in Fig. 2A. Its performance deteriorates as we increase c_{out}/c_{in} , with the disassortative one taking over with higher likelihood values. Furthermore, we can notice an inflexion point at $c_{out}/c_{in} = 1$, where the difference in likelihood between the models is null.

While one would expect the disassortative model to perform better in such a scenario, we highlight that this regime is a challenging and noisy one, as the affinity matrix is the uniform matrix of ones. Hence, recovery is difficult and not guaranteed, regardless of the model. We finally notice an increase of $\mathcal{L}_a - \mathcal{L}_d$ with c_{in} , which regulates the strength of the signal and makes it easier to separate the two regimes.

While we expect recovery to improve at more detectable regimes, this may not be observed by only looking at the $\mathcal{L}_a - \mathcal{L}_d$ difference. For this reason, in Fig. 2B, we complement our analysis by plotting only the log-likelihood \mathcal{L}_d attained via the disassortative initialization. In this case, we notice that the performance of the disassortative model increases with both c_{out}/c_{in} and c_{in} , as the intercommunity interactions get stronger and the expected degree gets higher. Altogether, our algorithm provides a principled way to extract arbitrary community interactions from higher-order data with varying structural organizations.

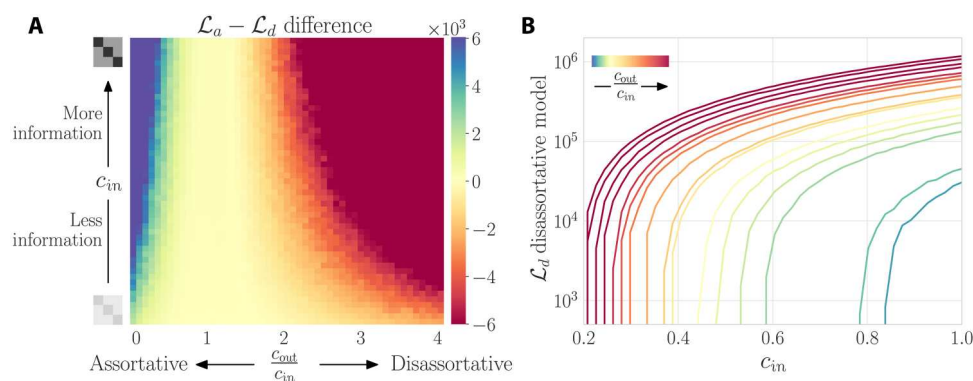


Fig. 2. Detection of assortative and disassortative community interactions. We generate data where the affinity matrices contain diagonal values c_{in} and out-diagonal c_{out} and measure the ability of our model to detect different assortative and disassortative regimes. (A) Positive (negative) differences in log-likelihood values indicate that the assortative (disassortative) model attains a better fit. An intermediate regime, highlighted in yellow, also emerges. Here, the detectability is compromised due to not having enough structure ($c_{out} \approx c_{in}$) or enough information (low c_{in}). (B) Log-likelihood of the disassortative model. In this case, the model attains better fit for data with marked disassortative structure (darker red).

CORE-PERIPHERY STRUCTURE

Many real-world systems are characterized by a different mesoscale organization known as core-periphery (CP) structure (59, 60). Networks characterized by such structure present a group of core of nodes connected among themselves, and often with high degree (61, 62), and a separate periphery of weakly connected nodes. Recently, methods to study and detect the existence of such patterns in hypergraphs have been proposed (63, 64). Conceptually, Hy-MMSBM has not been developed with the purpose of CP detection. Nevertheless, we can show its ability in capturing CP structures in hypergraphs through the generation of synthetic data that resemble the core structures of the input dataset.

To measure the recovery of CP structures, we use the method developed by Tudisco and Higham (64), HyperNSM, that assigns to each node of a hypergraph a core-score quantifying how close the node is to the core, where higher values denote stronger participation. HyperNSM achieved good performance on synthetic and real-world data, and its implementation is extremely efficient.

We analyze the Enron email dataset (65). Notably, the dataset comes with metadata information identifying a group of core nodes, employees of the organization who send batch emails to the periphery, which in turn only receive emails. This allows us to evaluate the ability of a model to recover a CP structure. In our study, we use the dataset used by Tudisco and Higham (64) with a planted core set that arises directly from the data collection process, as discussed by Amburg *et al.* (63) (it is preprocessed by keeping only hyperedges of size $D \leq 25$). The dataset has $N = 4423$ nodes and a core composed by 132 nodes. We apply HyperNSM to quantify the CP structure of the input Enron email dataset, as well as of the samples generated with Hy-MMSBM. To generate the samples, we first run our inference procedure on the Enron email dataset and then sample hypergraphs distributed according to the obtained u, w parameters. Further details on how to generate the samples are provided in Appendix Core-periphery experiments in the Supplementary Materials. For comparison, we also generate samples with a configuration model for hypergraphs (66) and obtain their core-score vectors with HyperNSM as well.

To evaluate the quality of the CP assignments in the different samples, we use the CP profile, the metric defined in (64) as

$$\gamma(S) = \frac{\# \text{ hyperedges with all nodes in } S}{\# \text{ hyperedges with at least one node in } S}, S \subseteq V \quad (10)$$

For any $k \in \{1, \dots, N\}$, we calculate the value $\gamma[S_k(x)]$, where $S_k(x)$ is the set of k nodes with smallest core-score in x . Given its definition, $\gamma(S)$ is small if S is largely contained in the periphery of the hypergraph and it should increase drastically as k crosses some threshold value k_0 , which indicates that the nodes in $\bigvee S_{k_0}(x)$ form the core.

In Fig. 3A we show the CP profiles corresponding to the core-scores computed with HyperNSM on the different datasets, i.e., the input Enron email, the samples generated with Hy-MMSBM, and the samples generated with the configuration model for hypergraphs. We plot 600 nodes with the highest core-score in decreasing order, and for all datasets, we notice a sharp drop, which highlights the existence of a CP structure. The main difference is given by the threshold k_0 at which this drop happens. This determines the dimension of the core. Remember that the data have a core composed

by 132 nodes, and when applying HyperNSM on the input data, we obtain a core dimension equal to 117, validating the good core-detection performance of this algorithm. The samples generated with the configuration model present a core with an average of 530.6 nodes, quite far from what observed in the input dataset. On the other hand, Hy-MMSBM generates samples that better resemble the property of the Enron email dataset, with an average core dimension of 195.7 nodes.

To understand the impact of nonpairwise interactions on higher-order CP structure, we also study the connection between hyperedge size and CP score. In Fig. 3B, we plot the CP score of a given node against the mean size of the hyperedges it belongs to. While we can observe a strong relationship between these two quantities at low CP scores, such regularity disappears in the center of the plot, which contains core nodes and presents a high scattering of hyperedge size values. This unexplained variance is justified by the rich information encoded in the CP score, which jointly depends on different factors related to the topology of the hypergraph. Yet, the scatter plots obtained on the Enron email dataset and the samples generated with Hy-MMSBM have higher similarity than the samples generated with the configuration model. Quantitatively, we measure the similarity between the core-scores of the different datasets for the 132 core nodes with the Pearson correlation, a measure $\rho \in [-1, 1]$ of linear correlation between two sets of data. The CP scores of the data have a Pearson correlation equal to 0.81 ± 0.01 with the samples generated with Hy-MMSBM, and of 0.76 ± 0.03 with the samples generated with the configuration model. Similar results are found on the relation between CP score and another structural property, namely, the degree of a node (see fig. S2 in Appendix Additional results on the Enron email dataset in the Supplementary Materials).

MODELING OF REAL DATA

In this section, we perform an extensive investigation of higher-order real-world systems. As explained in the "Inference" section and in the Supplementary Materials (Appendix Computational considerations), the linear-cost EM updates, together with a careful implementation that exploits the sparsity of most datasets, make our method suitable for the analysis of a variety of hypergraphs that were previously inaccessible due to computational constraints. Our method proves to be scalable with respect to both the number of system units and the size of the interactions, improving substantially on competing algorithms currently available in the literature. Moreover, our model is based on a probabilistic formulation, allowing it to perform additional operations and extract information that is not viable via other approaches, such as spectral clustering. First, we evaluate the quality of fit of various community detection methods based on their hyperedge prediction capabilities on a Gene Disease dataset, where nodes are genes, and interactions contain genes that are associated with a disease. To this end, we use the area under the curve (AUC) measure, a link prediction metric defined as follows: Given a randomly selected observed edge, and a randomly selected nonobserved one, the $AUC \in [0, 1]$ computes the number of times that the generative model assigns a higher probability to the observed edge. Here, we split the datasets into train and test subsets, where the train sets are used to estimate the parameters, and we evaluate the prediction performance in terms of AUC on the test sets (see Appendix Experiments on real data in the

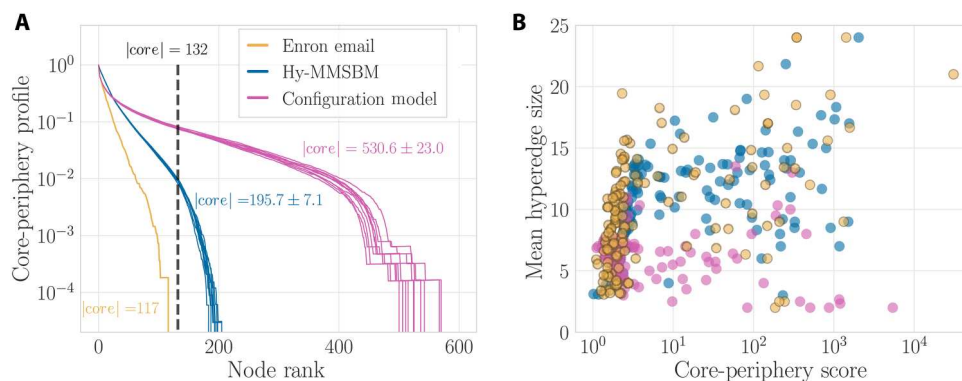


Fig. 3. Recovery of structural CP information. (A) CP profile (Eq. 10) corresponding to the core-scores computed with HyperNSM on the input Enron email (yellow), 10 synthetic samples generated with Hy-MMSBM (blue), and 10 synthetic samples generated with a configuration model for hypergraphs (magenta). We plot 600 nodes with the highest core-score in decreasing order and report the averages and standard deviations of the core dimension for the different datasets. Our method generates samples that closely resemble the property of the input dataset, with an average core dimension close to 132 nodes. (B) Mean size of the hyperedges a node belongs to against its CP score. We observe higher agreement between the data and the inference-based sample generated with Hy-MMSBM. This is also highlighted by the Pearson correlation of the 132 core nodes that is equal to 0.81 ± 0.01 for Hy-MMSBM versus the value of 0.76 ± 0.03 for the samples generated with the configuration model.

Supplementary Materials for details). Scalability with respect to hyperedge size is a crucial aspect of models for higher-order data. However, due to computational and numerical constraints, previous methods are limited to considering interactions of moderate size only, possibly causing a loss of information and a biased representation of the full system. In contrast, our model is able to efficiently process all the information provided in the dataset, reliably scaling to hyperedges of size of the order of the thousands. In Fig. 4A, we compare our method with other probabilistic approaches with hyperedge prediction capabilities. When only small interactions are considered, our model outperforms the competitive algorithms. At the computational limit of other approaches $D = 25$, Hypergraph-MT and our model attain a similar score, signaling the importance of considering large interactions. Beyond this computational threshold, our method continues to exploit the information provided by interactions among a growing number of units up to the maximum size observed of $D = 1074$, which results in an AUC score of 0.79.

We then extend our analysis to a variety of datasets from different domains, as described in Fig. 4B. For each dataset, we show the

inference running time as a function of the number of nodes N and the size of the largest hyperedge D . The AUC scores, reported in Table 1 and ranging from 0.74 to 0.98, show that the model generally yields a good fit and predicts the existence of hyperedges reliably. While these scores are on average aligned with those of other existing algorithms (35), the running time of our model is orders of magnitude lower. This allows studying very large hypergraphs such as the Arxiv, Trivago 2core, and Amazon datasets, containing up to millions of nodes and hyperedges. Overcoming the resulting computational challenges, our method allows the efficient modeling of a variety of previously unexplored datasets, which, to the best of our knowledge, could not be tackled by competing higher-order community detection algorithms.

Taken all together, these results show the effectiveness of our model in tackling datasets of small and large dimensions, in terms of both quantitative performance and computational scalability, and make Hy-MMSBM a valid tool for the study of complex higher-order systems.

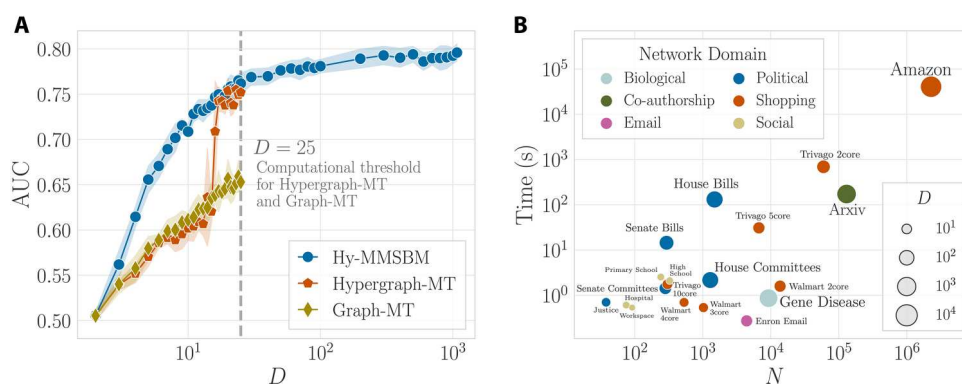


Fig. 4. Modeling of real data: hyperedge prediction and running time. (A) Quality of hyperedge prediction measured by the AUC score on a Gene Disease dataset, where nodes are genes and hyperedges contain genes that are associated with a disease. For Hypergraph-MT and Graph-MT, the plot shows a computational threshold at the maximum hyperedge size $D = 25$. Hy-MMSBM attains the highest scores and is able to model the entire hypergraph, up to $D = 1074$. (B) Running time of Hy-MMSBM for a variety of real-world datasets. The node represents the data domain. Both N and D are in log scale. The corresponding AUC scores are reported in Table 1.

DISCUSSION

Here, we have developed a probabilistic framework to model hypergraphs. Our method allows performing inference on very large hypergraphs, detecting their community structure, and reliably predicting the existence of higher-order interactions of arbitrary size. When compared to other available methods on synthetic hypergraphs with known ground truth, for both hard and mixed assignments, our model attains the most efficient recovery of the planted partitions. Moreover, compared to previous proposals, Hy-MMSBM relies on less restrictive assumptions on the latent

community structure in the data and is thus able to detect configurations, such as disassortative community interactions, which could not be previously identified. Furthermore, our method is extremely fast. Its efficient numerical implementation exploits optimized closed-form updates and dataset sparsity and has linear cost in the number of nodes and hyperedges. The resulting formulas are also numerically stable, not resulting in under- or overflows during the computations. Such numerical stability carries over to extremely large interactions, a substantial improvement over the computational threshold of previous methods, allowing to explore higher-order datasets with millions of nodes and interactions among thousands of units, that could not be previously tackled.

There are several directions for future work. From a theoretical perspective, our proposed likelihood function is based on a bilinear form for capturing dependencies within the hyperedges, a key ingredient for ensuring both mixed-membership nodes and fast inference. A possible extension would be to consider alternative likelihood definitions where the probability of the hyperedges is determined by multilinear forms, which would in principle allow capturing more complex interactions within the hyperedges. Similarly, here, we have assumed the hyperedges to be independent conditioned on the latent variables. Relaxing this assumption may ameliorate the expressiveness of the model, allowing to capture topological properties that involve more than two hyperedges, as already observed in the case of networks (67–69). From an algorithmic perspective, there are different questions that may allow further stabilizing and improving the inference procedure. Among these, the propensity of different initial conditions to be trapped in local optima during EM or MAP inference has not yet been investigated. Devising suitable initialization procedures or parameter priors to favor different membership types, as done in other works (70), offers a promising path in this direction. Finally, we have considered here a standard scenario where the input data are a list of hyperedges, and these are provided all at once. Other approaches may be needed in case of availability of extra information such as node attributes (71, 72) or for dynamic data (73).

Altogether, our work provides an accurate, flexible, and scalable tool for the modeling of very large hypergraphs, advancing our ability to tackle and study the organization of real-world higher-order systems.

Supplementary Materials

This PDF file includes:

Supplementary Text

Figs. S1 and S2

References

REFERENCES AND NOTES

1. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
2. F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J. G. Young, G. Petri, Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
3. L. Torres, A. S. Blevins, D. Bassett, T. Eliassi-Rad, The why, how, and when of representations for complex systems. *SIAM Rev.* **63**, 435–485 (2021).
4. F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, M. M. Murray, T. P. Peixoto, F. Vaccarino, G. Petri, The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
5. F. Battiston, G. Petri, *Higher-Order Systems* (Springer, 2022).

Table 1. AUC scores on real datasets. We report the number of nodes N , number of hyperedges $|E|$, maximum hyperedge size D , number of communities K , and AUC scores attained by our method on 19 large-scale real-world hypergraphs. The results are averages and standard deviations over 10 random test sets, and the value of K is chosen via cross-validation (see Appendix Experiments on real data in the Supplementary Materials).

	N	$ E $	D	K	AUC
Justice	38	2,826	9	4	0.909 ± 0.008
Hospital	75	1,825	5	2	0.767 ± 0.013
Workspace	92	788	4	5	0.741 ± 0.015
Primary School	242	12,704	5	10	0.832 ± 0.002
Senate Committees	282	301	31	30	0.926 ± 0.023
Senate Bills	294	21,721	99	13	0.921 ± 0.002
Trivago 10core	303	3,162	14	11	0.960 ± 0.005
High School	327	7,818	5	17	0.879 ± 0.007
Walmart 4core	532	2,292	10	4	0.837 ± 0.013
Walmart 3core	1,025	3,553	11	4	0.825 ± 0.010
House Committees	1,290	335	81	25	0.939 ± 0.015
House Bills	1,494	54,933	399	19	0.946 ± 0.001
Enron Email	4,423	5,734	25	2	0.835 ± 0.009
Trivago 5core	6,687	33,963	26	30	0.962 ± 0.001
Gene Disease	9,262	3,128	1,074	2	0.828 ± 0.010
Walmart 2core	13,706	19,869	25	2	0.788 ± 0.004
Trivago 2core	59,536	140,698	52	100	0.863 ± 0.002
Arxiv	130,024	172,173	2,097	10	0.884 ± 0.001
Amazon	2,268,231	4,242,421	9,350	29	0.978 ± 0.002

6. A. Patania, G. Petri, F. Vaccarino, The shape of collaborations. *EPJ Data Sci.* **6**, 18 (2017).
7. S. Klamt, U.-U. Haus, F. Theis, Hypergraphs and cellular networks. *PLOS Comput. Biol.* **5**, e1000385 (2009).
8. A. Zimmer, I. Katzir, E. Dekel, A. E. Mayo, U. Alon, Prediction of multidimensional drug dose responses based on measurements of drug pairs. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10442–10447 (2016).
9. G. Cencetti, F. Battiston, B. Lepri, M. Karsai, Temporal properties of higher-order interactions in social networks. *Sci. Rep.* **11**, 7028 (2021).
10. F. Musciotto, D. Papageorgiou, F. Battiston, D. R. Farine, Beyond the dyad: Uncovering higher-order structure within cohesive animal groups. bioRxiv 2022.05.30.494018 [Preprint]. 30 May 2022. <https://doi.org/10.1101/2022.05.30.494018>.
11. G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, F. Vaccarino, Homological scaffolds of brain functional networks. *J. R. Soc. Interface* **11**, 20140873 (2014).
12. C. Giusti, R. Ghrist, D. S. Bassett, Two's company, three (or more) is a simplex. *J. Comput. Neurosci.* **41**, 1–14 (2016).
13. A. Santoro, F. Battiston, G. Petri, E. Amico, Higher-order organization of multivariate time series. *Nat. Phys.* **19**, 1–9 (2023).
14. T. Carletti, F. Battiston, G. Cencetti, D. Fanelli, Random walks on hypergraphs. *Phys. Rev. E* **101**, 022308 (2020).
15. C. Bick, P. Ashwin, A. Rodrigues, Chaos in generically coupled phase oscillator networks with nonpairwise interactions. *J. Nonlin. Sci.* **26**, 094814 (2016).
16. P. S. Skardal, A. Arenas, Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.* **3**, 218 (2020).
17. A. P. Millán, J. J. Torres, G. Bianconi, Explosive higher-order kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.* **124**, 218301 (2020).
18. M. Lucas, G. Cencetti, F. Battiston, Multiorder laplacian for synchronization in higher-order networks. *Phys. Rev. Res.* **2**, 033410 (2020).
19. L. V. Gambuzza, F. Di Patti, L. Gallo, S. Lepri, M. Romance, R. Criado, M. Frasca, V. Latora, S. Boccaletti, Stability of synchronization in simplicial complexes. *Nat. Commun.* **12**, 1255 (2021).
20. Y. Zhang, M. Lucas, F. Battiston, Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. *Nat. Commun.* **14**, 1605 (2023).
21. I. Iacopini, G. Petri, A. Barrat, V. Latora, Simplicial models of social contagion. *Nat. Commun.* **10**, 2485 (2019).
22. S. Chowdhary, A. Kumar, G. Cencetti, I. Iacopini, F. Battiston, Simplicial contagion in temporal higher-order networks. *J. Phys. Complex.* **2**, 035019 (2021).
23. L. Neuhäuser, A. Mellor, R. Lambiotte, Multibody interactions and nonlinear consensus dynamics on networked systems. *Phys. Rev. E* **101**, 032310 (2020).
24. U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, V. Latora, Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.* **5**, 586–595 (2021).
25. A. Civilini, N. Anbarci, V. Latora, Evolutionary game model of group choice dilemmas on hypergraphs. *Phys. Rev. Lett.* **127**, 268301 (2021).
26. A. Civilini, O. Sadekar, F. Battiston, J. Gómez-Gardeñes, V. Latora, Explosive cooperation in social dilemmas on higher-order networks. arXiv:2303.11475 [physics.soc-ph] (20 March 2023).
27. C. Berge, *Graphs and Hypergraphs* (North-Holland Pub. Co., 1973).
28. A. R. Benson, Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.* **1**, 293–312 (2019).
29. F. Tudisco, D. J. Higham, Node and edge nonlinear eigenvector centrality for hypergraphs. *Commun. Phys.* **4**, 201 (2021).
30. A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11221–E11230 (2018).
31. Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, Higher-order motif analysis in hypergraphs. *Commun. Phys.* **5**, 79 (2022).
32. Q. F. Lotito, F. Musciotto, F. Battiston, A. Montresor, Exact and sampling methods for mining higher-order motifs in large hypergraphs. arXiv:2209.10241 [cs.SI] (21 September 2022).
33. F. Musciotto, F. Battiston, R. N. Mantegna, Detecting informative higher-order interactions in statistically validated hypergraphs. *Commun. Phys.* **4**, 218 (2021).
34. F. Musciotto, F. Battiston, R. N. Mantegna, Identifying maximal sets of significantly interacting nodes in higher-order networks. arXiv:2209.12712 [physics.soc-ph] (26 September 2022).
35. M. Contisciani, F. Battiston, C. De Bacco, Inference of hyperedges and overlapping communities in hypergraphs. *Nat. Commun.* **13**, 7229 (2022).
36. J.-G. Young, G. Petri, T. P. Peixoto, Hypergraph reconstruction from network data. *Commun. Phys.* **4**, 135 (2021).
37. K. Balasubramanian, D. Gitelman, H. Liu, Nonparametric modeling of higher-order interactions via hypergraphons. *J. Mach. Learn. Res.* **22**, 146 (2021).
38. Z. T. Ke, F. Shi, D. Xia, Community detection for hypergraph networks via regularized tensor power iteration. arXiv:1909.06503 [stat.ME] (14 September 2019).
39. K. Turnbull, S. Lunagomez, C. Nemeth, E. Airolidi, Latent space modelling of hypergraph data. arXiv:1909.00472 [stat.ME] (1 September 2019).
40. T. L. J. Ng, T. B. Murphy, Model-based clustering for random hypergraphs. *Adv. Data Anal. Classif.* **16**, 691–723 (2022).
41. T. Carletti, D. Fanelli, R. Lambiotte, Random walks and community detection in hypergraphs. *J. Phys. Complex.* **2**, 015011 (2021).
42. A. Eriksson, D. Edler, A. Rojas, M. de Domenico, M. Rosvall, How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun. Phys.* **4**, 133 (2021).
43. D. Zhou, J. Huang, B. Schölkopf, Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **19**, 1601–1608 (2006).
44. D. Ghoshdastidar, A. Dukkipati, A provable generalized tensor spectral method for uniform hypergraph partitioning, in *International Conference on Machine Learning* (PMLR, 2015), pp. 400–409.
45. M. C. Angelini, F. Caltagirone, F. Krzakala, L. Zdeborová, Spectral detection on sparse hypergraphs, in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing* (Allerton) (IEEE, 2015), pp. 66–73.
46. X. Gong, D. J. Higham, K. Zygalakis, Generative hypergraph models and spectral embedding. *Sci. Rep.* **13**, 540 (2023).
47. D. Ghoshdastidar, A. Dukkipati, Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Adv. Neural Inf. Process. Syst.* **27**, (2014).
48. C.-Y. Lin, I. E. Chien, L.-H. Wang, On the fundamental statistical limit of community detection in random hypergraphs, in *2017 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2017), pp. 2178–2182.
49. K. Ahn, K. Lee, C. Suh, Community recovery in hypergraphs. *IEEE Trans. Inf. Theory* **65**, 6561–6579 (2019).
50. P. S. Chodrow, N. Veldt, A. R. Benson, Generative hypergraph clustering: From blockmodels to modularity. *Sci. Adv.* **7**, eabh1303 (2021).
51. L. Brusa, C. Matias, Model-based clustering in simple hypergraphs through a stochastic blockmodel. arXiv:2210.05983 [stat.ME] (12 October 2022).
52. N. Ruggeri, F. Battiston, C. De Bacco, A framework to generate hypergraphs with community structure. arXiv:2212.08593 [cs.SI] (22 June 2023).
53. E. M. Airolidi, D. Blei, S. Fienberg, E. Xing, Mixed membership stochastic blockmodels. *Adv. Neural Inf. Process. Syst.* **9**, 1981–2014 (2008).
54. C. De Bacco, E. A. Power, D. B. Larremore, C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
55. Q. F. Lotito, M. Contisciani, C. de Bacco, L. di Gaetano, L. Gallo, A. Montresor, F. Musciotto, N. Ruggeri, F. Battiston, HypergraphX: A library for higher-order network analysis. *Journal of Complex Networks* **11**, cnad019 (2023).
56. N. Veldt, A. R. Benson, J. Kleinberg, Combinatorial characterizations and impossibilities for higher-order homophily. *Sci. Adv.* **9**, eabq3200 (2023).
57. L. Peel, D. B. Larremore, A. Clauset, The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
58. A. Decelle, F. Krzakala, C. Moore, L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).
59. S. P. Borgatti, M. G. Everett, Models of core/periphery structures. *Soc. Netw.* **21**, 375–395 (2000).
60. P. Csermely, A. London, L.-Y. Wu, B. Uzzi, Structure and dynamics of core/periphery networks. *J. Complex Netw.* **1**, 93–123 (2013).
61. V. Colizza, A. Flammini, M. A. Serrano, A. Vespignani, Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).
62. A. Ma, R. J. Mondragón, Rich-cores in networks. *PLOS ONE* **10**, e0119678 (2015).
63. I. Amburg, J. Kleinberg, A. R. Benson, Planted hitting set recovery in hypergraphs. *J. Phys. Complex* **2**, 035004 (2021).
64. F. Tudisco, D. J. Higham, Core-periphery detection in hypergraphs. *SIAM J. Math. Data Sci.* **5**, 1–21 (2023).
65. B. Klimt, Y. Yang, *European Conference on Machine Learning* (Springer, 2004), pp. 217–226.
66. P. S. Chodrow, Configuration models of random hypergraphs. *Networks* **8**, cnaa018 (2020).
67. H. Safdari, M. Contisciani, C. De Bacco, Generative model for reciprocity and community detection in networks. *Phys. Rev. Res.* **3**, 023209 (2021).
68. M. Contisciani, H. Safdari, C. De Bacco, Community detection and reciprocity in networks by jointly modelling pairs of edges. *Networks* **10**, cna034 (2022).
69. H. Safdari, M. Contisciani, C. De Bacco, Reciprocity, community detection, and link prediction in dynamic networks. *J. Phys. Complex* **3**, 015010 (2022).

70. N. Nakis, A. Çelikkanat, M. Mørup, *Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and Their Applications: COMPLEX NETWORKS 2022–Volume 1* (Springer, 2023), pp. 350–363.
71. M. E. Newman, A. Clauset, Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).
72. M. Contisciani, E. A. Power, C. De Bacco, Community detection with node attributes in multilayer networks. *Sci. Rep.* **10**, 15736 (2020).
73. C. Matias, V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 1119–1141 (2017).
74. E. L. Lehmann, G. Casella, *Theory of Point Estimation* (Springer Science & Business Media, 2006).
75. D. B. Larremore, A. Clauset, A. Z. Jacobs, Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90**, 012805 (2014).
76. N. W. Landry, M. Lucas, I. Iacopini, G. Petri, A. Schwarze, A. Patania, L. Torres, Xgi: A python package for higher-order interaction networks. *J. Open Source Softw.* **8**, 5162 (2023).

Acknowledgments

Funding: N.R. acknowledges support from the Max Planck ETH Center for Learning Systems. M.C. and C.D.B. were supported by the Cyber Valley Research Fund. M.C. acknowledges support from the International Max Planck Research School for Intelligent Systems (IMPRS-IS). F.B. acknowledges support from the Air Force Office of Scientific Research under award number FA8655-22-1-7025. **Author contributions:** All authors conceived the project. N.R. developed the code implementation and performed the simulations and analysis. All the authors contributed to the development of models and experiments and to the writing and revision of the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All synthetic data needed to evaluate the conclusions of the paper are explained in detail for reproduction. All real data are properly referenced and publicly available.

Submitted 30 January 2023

Accepted 12 June 2023

Published 12 July 2023

10.1126/sciadv.adg9159

Hypergraphx: a library for higher-order network analysis

QUINTINO FRANCESCO LOTITO

*Department of Information Engineering and Computer Science, University of Trento, via Sommarive 9,
38123 Trento, Italy*

†Corresponding author. Email: quintino.lotito@unitn.it

MARTINA CONTISCIANI, CATERINA DE BACCO

Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany

LEONARDO DI GAETANO, LUCA GALLO

Department of Network and Data Science, Central European University, 1100 Vienna, Austria

ALBERTO MONTRESOR

*Department of Information Engineering and Computer Science, University of Trento, via Sommarive 9,
38123 Trento, Italy*

FEDERICO MUSCIOTTO

*Dipartimento di Fisica e Chimica Emilio Segrè, Università di Palermo, Viale delle Scienze, Ed. 18,
I-90128, Palermo, Italy*

NICOLÒ RUGGERI

*Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany and
Department of Computer Science, ETH, 8004 Zürich, Switzerland*

AND

FEDERICO BATTISTON

Department of Network and Data Science, Central European University, 1100 Vienna, Austria

[Received on 29 March 2023; editorial decision on 28 April 2023; accepted on 10 May 2023]

From social to biological systems, many real-world systems are characterized by higher-order, non-dyadic interactions. Such systems are conveniently described by hypergraphs, where hyperedges encode interactions among an arbitrary number of units. Here, we present an open-source python library, hypergraphx (HGX), providing a comprehensive collection of algorithms and functions for the analysis of higher-order networks. These include different ways to convert data across distinct higher-order representations, a large variety of measures of higher-order organization at the local and the mesoscale, statistical filters to sparsify higher-order data, a wide array of static and dynamic generative models, and an implementation of different dynamical processes with higher-order interactions. Our computational framework is general, and allows to analyse hypergraphs with weighted, directed, signed, temporal and multiplex group interactions. We provide visual insights on higher-order data through a variety of different visualization tools. We accompany our code with an extended higher-order data repository and demonstrate the ability of HGX to analyse real-world systems through a systematic analysis of a social network with higher-order interactions. The library is conceived as an evolving, community-based effort, which will further extend its functionalities over the years. Our software is available at <https://github.com/HGX-Team/hypergraphx>.

Keywords: higher-order networks; hypergraphs; complex networks; network analysis.

1. Introduction

In the last few decades, networks have emerged as the natural tool to model a wide variety of natural, social and man-made systems. Networks, collections of nodes and links connecting pairs of them, are able to capture dyadic interactions only. However, in many real-world systems units interact in groups of three or more [1–4]. Systems with non-dyadic interactions are ubiquitous, with examples ranging from cellular networks [5], drug recombination [6], structural and functional brain networks [7–9], human [10] and animal [11] face-to-face interactions, and collaboration networks [12]. These higher-order interactions can be naturally described by alternative mathematical structures such as hypergraphs [2, 13], where hyperedges connect groups of nodes of arbitrary size.

In the last 25 years, advances in technology have generated an unprecedented amount of relational data across a variety of domains. Broadening the scopes of the first pioneering contributions to the field of network science [14–16], these allowed to develop new data-informed frameworks to investigate biological, technological and social systems. In parallel with theoretical and methodological progresses, a crucial role in advancing network science has been played by the development of efficient algorithms and computational tools to analyse networked data. Nowadays, widely used, community-based software such as NetworkX [17] and igraph [18], and individual efforts such as graph-tool [19]—just to mention a few—have enabled thousands of researchers to perform multi-faceted, large-scale network analysis of relational data. Only recently, some early contributions [20–24], in particular XGI [25], have started to develop computational tools to match the explosion of interest in higher-order systems.

Here, we provide our contribution by presenting hypergraphx (HGX), a multi-purpose, open-source Python library for the analysis of networked systems with higher-order interactions. The library is conceived by researchers with several years of experience and direct contributions to the field of higher-order interactions. Developed by a diverse multidisciplinary team with complementary skills and expertise, HGX aims to provide, as a single source, a comprehensive suite of tools and algorithms for constructing, storing, analysing and visualizing systems with higher-order interactions. These include different ways to convert data across distinct higher-order representations, a large variety of measures of higher-order organization at the local and the mesoscale, statistical filters to sparsify higher-order data, a wide array of static and dynamic generative models, an implementation of different dynamical processes with higher-order interactions, from epidemics to diffusion and synchronization and more. Our computational framework is general, and allows to analyse hypergraphs with weighted, directed, signed, temporal and multiplex group interactions. Beyond experts in the field, we hope that our library will make higher-order network analysis accessible to everyone interested in exploring the higher-order dimension of relational data.

2. Tools

Here, we discuss the main functionalities provided by HGX. The different tools of our library are illustrated online through detailed, user-friendly tutorials. The library is conceived as an evolving, community-based effort, which will further extend its functionalities over the years.

2.1 Representations

Hypergraphs represent the most general and flexible framework to encode systems with higher-order interactions [2, 13]. However, specific research questions or data features might benefit from alternative higher-order frameworks. We provide functions to easily and efficiently convert higher-order data usually

represented as hypergraphs into different representations [2, 26] such as bipartite networks, maximal simplicial complexes, higher-order line graphs, dual hypergraphs and clique-expansion graphs.

2.2 Basic node and hyperedge statistics

Our library provides simple tools characterizing basic node and hyperedge statistics. These include measures of hyperdegree distributions, both aggregated or separated by order of interactions, as well as measures of correlations among them. We include functions to compute hyperdegree–hyperdegree assortativity, both within and across orders. We provide simple tools to compute hyperedge size distribution in the whole system, as well as at the level of individual nodes.

2.3 Centrality measures

Centrality scores are a key tool in network analysis, and allow to quantify the importance or influence of different nodes within a system [15]. Nodes with high centrality usually have a high number of links, are strategically connected to other influential nodes, or are characterized by both such features. Our library provides a variety of higher-order centrality measures, where interactions in different group sizes are taken into account. These include centrality measures based on node participation in different sub-hypergraphs [27] and different centrality scores based on spectral approaches [28]. We also implement measures of hyperedge centrality based on shortest paths and betweenness flows [29].

2.4 Motifs

Motifs are small recurring patterns of subgraphs that are over-represented in a network [30]. Motif analysis has established itself as a fundamental tool in network science to describe networked systems at their microscale, identifying their structural and functional building blocks [31]. We provide an implementation for higher-order motif analysis, extracting overabundant subgraphs of nodes connected by higher-order interactions, as originally defined in Ref. [32]. Given their widespread applications and expected use on large-scale real-world datasets, we also provide an approximated algorithm for higher-order motif analysis based on hyperedge sampling, able to speed up computations by orders of magnitudes with only a minimal compromise in accuracy [33].

2.5 Mesoscale structures

One of the most relevant features of graphs representing real-world systems is community structure [34]. A variety of approaches for community detection on graphs show how these partitions provide meaningful insights into the fundamental patterns underlying node interactions. Recently, methods for defining the mesoscale structure of higher-order networks have been explored. Here, we provide an implementation of a spectral method which recovers hard communities via hypergraph cut optimization [35]. We also implement different generative models able to extract overlapping communities and jointly infer hyperedges [36], allowing to capture a variety of mesoscale organizations, including both disassortative and assortative community structure [37]. We provide a method able to extract hyperlink communities, where interactions, and not system units, are clustered across different hypergraph modules [38]. Finally, we provide a method to extract the core–periphery organization of higher-order systems, capturing a group of central and tightly connected nodes in hypergraphs governing the overall system behaviour, inspired by Ref. [39].

2.6 Filters

Many real-world systems are characterized by an abundance of noisy and redundant interactions, resulting in overly densely connected networks. Different filtering techniques have been developed to identify the most informative links by adopting an approach based on statistical validation, where the statistical significance of interactions of the real system is evaluated by comparing them with an ensemble of random replicas that preserve some individual features (like degree or strength) [40]. Our library provides a variety of different tools to filter systems with higher-order interactions. These include extracting statistically validated hypergraphs, which are a collection of hyperlinks that are over-expressed representing nodes that are significantly interacting in the same exact group of fixed size [41] and identifying significant maximally interacting sets, which represent the largest groups of nodes that interact significantly, captured by combining interactions of different orders [42].

2.7 Generative models

The ability to produce synthetic data with different topological characteristics has proven crucial for a variety of tasks, from algorithms benchmarking to the study and testing of non-trivial network statistics [43, 44]. In our library, we offer ready-to-use implementations for various synthetic hypergraph samplers. We provide functions to build generalized Erdős–Rényi models, both for uniform (all interactions have the same order) and non-uniform (different orders of interactions) hypergraphs. We implement scale-free random hypergraph models with the possibility of tuning the correlation between the degree sequence among different orders. We also include a variety of randomization tools and a configuration model for hypergraphs, where samples are produced respecting given node degree and hyperedge size sequences [45]. Based on a similar mechanism, we implement also a more complex sampler which allows to specify hard and soft community assignments for nodes, and arbitrary community structure, such as assortative and disassortative [46]. Finally, we provide a higher-order activity-driven model with group interactions that change in time [47] and compute the associated percolation threshold.

2.8 Dynamical processes

The structural properties of complex networks shape the dynamical process occurring on top of them [48]. Recent works have revealed that higher-order interactions significantly impact various dynamical processes, including percolation [49], diffusion [50, 51], pattern formation [52, 53], synchronization [54–58], contagion [59–61] and evolutionary games [62–64]. We provide functions to investigate several of these processes. These include tools to study synchronization with higher-order interactions, from the analysis of the multiorder Laplacian matrix for kuramoto dynamics [55], to the implementation of the Master Stability Function approach for synchronization stability [54, 65]. We also provide an algorithm to simulate simplicial social contagion [60], and analytical and numerical tools to investigate random walks on hypergraphs [50].

2.9 Weighted, directed, signed, temporal and multiplex hypergraphs

Our library is highly flexible. It allows to store and analyse hypergraphs with a rich set of features associated with hyperedges, including interactions of different intensity, directions, sign, that vary in time or belong to different layers of a multiplex system.

2.10 Visualization

The adoption of higher-order networks is rapidly increasing, and the development of standard tools to visualize them is still in progress. Our library provides different visualization tools to gain visual insights into the higher-order organization of real-world systems. We provide tools to plot systems with higher-order interactions, where hyperedges of arbitrary size encode relationships among an arbitrary number of nodes. Due to the rapid combinatorial increase in the number of possible higher-order interactions and their overlaps, such a direct approach is particularly suited for systems with a moderate number of nodes, while such a visualization might not be effective in other cases. Therefore, we provide alternative solutions that may assist the practitioner in a variety of cases, such as relational data with a large number of nodes or large hyperedges. For instance, we give the option to plot the bipartite projection of a hypergraph where the two sets of nodes represent respectively the original system units and the hyperedges in which they take part. We can also plot the hypergraph clique projection, which results in a simple graph where each hyperedge of size s is decomposed into a clique of $\frac{s(s-1)}{2}$ unordered pairwise interactions. Additionally, we implement a multilayer representation of the hypergraph where each layer encodes interactions of a given size, and two nodes are connected in layer s only if they interact in the hypergraph through a hyperedge of size s . Finally, we offer a novel way of visualizing hypergraphs, where the hypergraph is represented as a graph whose nodes are pie charts. These pie charts indicate the proportion of interaction sizes for each node, and two nodes are connected when they have significant interactions across multiple orders.

3. Data

Here, we present the dataset repository accompanying our library. Such a repository is intended to provide an initial core of higher-order relational data, that we aim to expand over the next few years. We illustrate the functionalities of HGX by performing different higher-order analyses for one of these datasets.

3.1 Higher-order data repository

The availability of data plays a fundamental role in developing theoretical frameworks and computational tools across different scientific domains and applications. The recent explosion of higher-order relational data has led to novel methodologies to study higher-order systems, which in turn require extensive datasets to be tested and validated. A few of these data are inherently higher order. Several others, instead, have originally been investigated with pairwise approaches, but have recently been re-explored under the new lens of higher-order network analysis. This motivates us to accompany our library with an easily accessible and well-curated data repository, functioning as a unifying source of datasets for the analysis of higher-order systems. We provide a collection of datasets for higher-order systems across different domains, including ecological (animal proximity [66]), social (human face-to-face interactions [67–71], co-authorships [72–74], votes [75]), technological (e-mails [73, 76, 77]) and biological (gene-disease [78] and drug [73] associations) systems. Some of these datasets record metadata characterizing the system units (e.g. whether an individual in a hospital is a patient or a doctor) and the interactions among them (e.g. the scientific domain of a research paper involving a group of authors). Also, they store information about the structural features of group interactions, which can be non-reciprocal, multi-relational and time-varying. Datasets can be loaded to explicitly highlight some of these characteristics. Indeed, our library allows to apply filters in the data loading process, for example, by selecting specific sets of nodes with regard to some metadata restriction, or by extracting group interactions limited to a given size, type

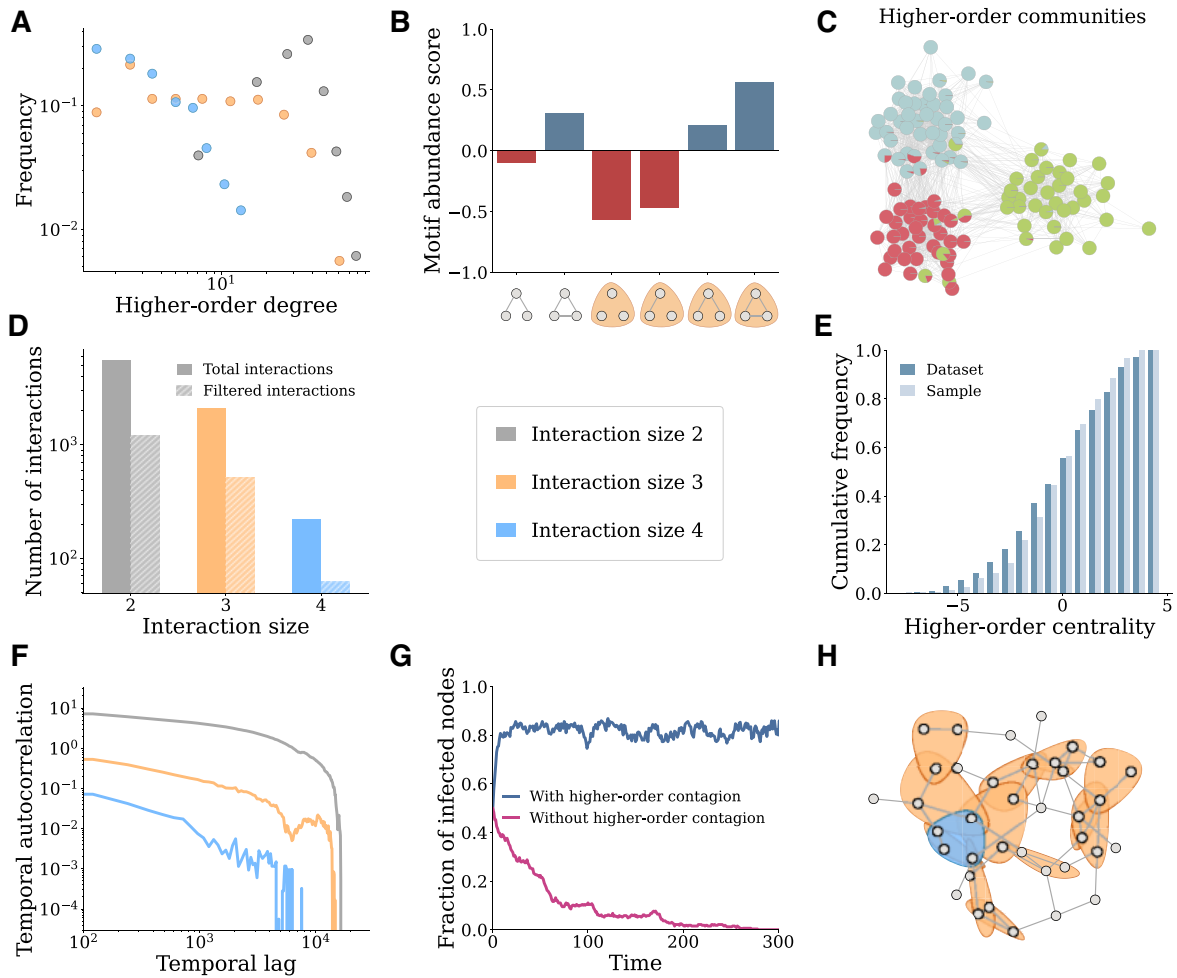


FIG. 1. Higher-order analysis of social interactions. We illustrate different functionalities of HGX on a dataset of face-to-face group interactions in a school from the SocioPattern collaboration [69]. (A) Higher-order degree distributions for different interaction sizes. (B) Higher-order motif analysis. (C) Higher-order overlapping community detection, and comparison with node metadata (we plot a subset of three classes). (D) Statistics of original and filtered higher-order social interactions. (E) Higher-order centrality measure in the dataset, and in sample obtained from a higher-order generative model. (F) Temporal autocorrelation for different sizes. (G) Fraction of infected nodes over time for a spreading process with or without higher-order infections. (H) Direct hypergraph visualization of social interactions (we plot a subset of one class, considering only statistically significant interactions).

or time interval. In the next years, we plan to continuously expand the data repository, and to add further filtering options to the data loading functions.

3.2 Analysing real-world higher-order systems: a guided tour

To illustrate the power of HGX in loading, manipulating, analysing and visualizing real-world systems with group interactions, in Fig. 1, we present an illustrative analysis of a dataset from the SocioPattern collaboration encoding face-to-face social interactions in a high school [69]. This dataset has been widely investigated in the literature on higher-order interactions [32, 36, 46, 60, 73], and records the activity of 327 students, divided into nine different high school classes. Our analysis focuses in particular on interactions among 2, 3 and 4 individuals, as statistics is limited for larger groups.

In Fig. 1(A), we show the different higher-order degree distributions. The largest degrees are obtained for pairwise interactions, and, in general, the curves show different profiles. Higher-order degree distributions display different correlations across different orders (Pearson’s correlation coefficient ρ , $\rho^{2,3} = 0.74$, $\rho^{2,4} = 0.46$, $\rho^{3,4} = 0.72$). To characterize such a higher-order system at the microscale, in Fig. 1(B), we perform higher-order motif analysis as introduced in Ref. [32]. We consider subhypergraphs of three nodes and capture over- (positive abundance score greater) and under- (negative) represented motifs in the data, as compared to a randomized higher-order configuration model [45]. Local structures with group interactions supported by pairwise links are found to be particularly relevant. In Fig. 1(C), we describe the mesoscale structure of the system, by extracting overlapping communities with the method of Ref. [36]. For simplicity, we consider a subset of three classes and plot pairwise interactions only. Nodes are represented as pie-charts, colored proportionally to the higher-order communities they belong to. In general, the inferred modules are well aligned with node metadata, with most students largely interacting within the community associated with their class. In Fig. 1(D), we show statistics for the interactions in the dataset. We see an inverse trend between the number of interactions and group size. We also plot statistics for a filtered system, where we have considered statistically validated hypergraphs [41], removing redundant hyperedges and identifying the most informative group interactions. We continue by showcasing the ability of the model introduced in Ref. [46] to generate hypergraphs which are similar to the original dataset. To validate such a statement, in Fig. 1(E), we plot the distribution of (a rescaled version of) higher-order centrality measure [27] both in the real and sampled hypergraphs, showing good agreement between the two. To further illustrate the flexibility of our computational framework, we then consider the temporal dimension of higher-order interactions. In particular, in Fig. 1(F), we show the temporal autocorrelation for different interaction sizes, one of the measures introduced to characterize the temporal evolution of higher-order systems in Ref. [79]. Results show the existence of long-range correlations at all orders of interactions, with a temporal cut-off which is dependent on the group size. Beyond structural analysis, our library also allows to investigate a variety of dynamical processes with higher-order interactions. Here, we simulate higher-order spreading among students in high school, following a model where groups of infected individuals are associated with higher-order contagion terms, in addition to traditional pairwise mechanisms [60]. In Fig. 1(G), we show the fraction of infected nodes over time for two configurations, one with and one without higher-order infections. As shown, the presence of such a higher-order term might significantly change the collective dynamics, pushing the system from a healthy to an endemic state. Finally, in Fig. 1(H), we present a direct hypergraph visualization of the higher-order system. For simplicity, we plot individuals belonging to a single class and display all statistically significant interactions [41] among two, three and four of them.

4. Conclusions

Hand in hand with new theory and methodologies, the development of efficient algorithms and software to analyse networked data has played a pivotal role in the advancement of modern network science. Here, we have presented HGX, a versatile and robust python library that offers a flexible and efficient framework to analyse networked systems with higher-order interactions. Its user-friendly environment and its vast range of functionalities make it accessible and useful to practitioners and researchers to answer a wide variety of needs and questions. In the future, we aim to keep expanding the toolkit of HGX across multiple new dimensions. For instance, we can already foresee the implementation of tools to investigate the robustness of higher-order systems under different attack strategies. We will also provide methods to efficiently summarize higher-order information and reduce the dimensionality of higher-order data. We aim to include tools to build and analyse higher-order dependencies from multivariate time series [9],

and measures of information theory to capture redundant and synergistic higher-order interactions [80]. Moreover, we aim to expand our coverage of higher-order processes, by including different evolutionary games [62, 64], ecological dynamics [81] and more.

We hope that HGX will make higher-order network analysis open to all researchers dealing with networked data, and we invite the community to explore the library and contribute.

Funding

The Air Force Office of Scientific Research under award number (FA8655-22-1-7025 to F.B. and L.G.; European Union through Horizon Europe CLOUDSTARS project (101086248); The International Max Planck Research School for Intelligent Systems (IMPRS-IS) (to M.C.); the Cyber Valley Research Fund (to C.D.B. and M.C.); the Max Planck ETH Center for Learning Systems (to N.R.).

Author contributions

F.B. and Q.F.L. coordinated the project. Q.F.L. is the leading developer. All authors contributed tools to the library, wrote and revised the article.

REFERENCES

1. BATTISTON, F., AMICO, E., BARRAT, A., BIANCONI, G., FERRAZ DE ARRUDA, G., FRANCESCHIETTO, B., IACOPINI, I., KÉFI, S., LATORA, V., MORENO, Y. ET AL. (2021) The physics of higher-order interactions in complex systems. *Nat. Phys.*, **17**, 1093–1098.
2. BATTISTON, F., CENCETTI, G., IACOPINI, I., LATORA, V., LUCAS, M., PATANIA, A., YOUNG, J.-G. & PETRI, G. (2020) Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.*, **874**, 1–92.
3. BATTISTON, F. & PETRI, G. (2022) *Higher-Order Systems*. Springer.
4. BIANCONI, G. (2021) *Higher-Order Networks*. Cambridge University Press.
5. KLAMT, S., HAUS, U.-U. & THEIS, F. (2009) Hypergraphs and cellular networks. *PLoS Comput. Bio.*, **5**, e1000385.
6. ZIMMER, A., KATZIR, I., DEKEL, E., MAYO, A. E. & ALON, U. (2016) Prediction of multidimensional drug dose responses based on measurements of drug pairs. *Proc. Natl. Acad. Sci., USA* **113**, 10442–10447.
7. GIUSTI, C., GHRIST, R. & BASSETT, D. S. (2016) Two's company, three (or more) is a simplex. *J. Comput. Neurosci.*, **41**, 1–14.
8. PETRI, G., EXPERT, P., TURKHEIMER, F., CARHART-HARRIS, R., NUTT, D., HELLYER, P. J. & VACCARINO, F. (2014) Homological scaffolds of brain functional networks. *J. R. Soc. Interface*, **11**, 20140873.
9. SANTORO, A., BATTISTON, F., PETRI, G. & AMICO, E. (2023) Higher-order organization of multivariate time series. *Nat. Phys.*, 1–9.
10. CENCETTI, G., BATTISTON, F., LEPRI, B. & KARSAI, M. (2021) Temporal properties of higher-order interactions in social networks. *Sci. Rep.*, **11**, 1–10.
11. MUSCIOTTO, F., PAPAGEORGIOU, D., BATTISTON, F. & FARINE, D. R. (2022) Beyond the dyad: uncovering higher-order structure within cohesive animal groups. bioRxiv, not peer reviewed.
12. PATANIA, A., PETRI, G. & VACCARINO, F. (2017) The shape of collaborations. *EPJ Data Sci.*, **6**, 1–16.
13. BERGE, C. (1973) *Graphs and hypergraphs*. North-Holland Pub. Co. USA
14. BARABÁSI, A.-L. & ALBERT, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.
15. WASSERMAN, S. & FAUST, K. (1994) *Social network analysis: methods and applications*.
16. WATTS, D. J. & STROGATZ, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
17. HAGBERG, A., SWART, P. & SCHULT, D. (2008) Exploring network structure, dynamics, and function using NetworkX. Technical Report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

18. CSARDI, G., NEPUSZ, T. ET AL. (2006) The igraph software package for complex network research. *InterJ. Complex Sys.*, **1695**, 1–9.
19. PEIXOTO, T. P. (2014) The graph-tool python library. *figshare*.
20. (2021) HyperNetworkX
21. ANTELMÍ, A., CORDASCO, G., KAMIŃSKI, B., PRAŁAT, P., SCARANO, V., SPAGNUOLO, C. & SZUFEL, P. (2020) Analyzing, exploring, and visualizing complex networks via hypergraphs using SimpleHypergraphs. *jl. arXiv, arXiv:2202.04654*, preprint, not peer reviewed.
22. BADIE-MODIRI, A. & KIVELÄ, M. (2022) Reticula: a temporal network and hypergraph analysis software package
23. DIAZ, L. P. & STUMPF, M. P. (2022) HyperGraphs. *jl: representing higher-order relationships in Julia. Bioinformatics*, **38**, 3660–3661.
24. MARCHETTE, D. J. (2021) HyperG
25. LANDRY, N., TORRES, L., IACOPINI, I., LUCAS, M., PETRI, G., PATANIA, A. & SCHWARZE, A. (2022) XGI
26. TORRES, L., BLEVINS, A. S., BASSETT, D. & ELIASSI-RAD, T. (2021) The why, how, and when of representations for complex systems. *SIAM Rev.*, **63**, 435–485.
27. ESTRADA, E. & RODRÍGUEZ-VELÁZQUEZ, J. A. (2006) Subgraph centrality and clustering in complex hyper-networks. *Physica A*, **364**, 581–594.
28. BENSON, A. R. (2019) Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.*, **1**, 293–312.
29. AKSOY, S. G., JOSLYN, C., MARRERO, C. O., PRAGGASTIS, B. & PURVINE, E. (2020) Hypernetwork science via high-order hypergraph walks. *EPJ Data Sci.*, **9**, 16.
30. MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. & ALON, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
31. MILO, R., ITZKOVITZ, S., KASHTAN, N., LEVITT, R., SHEN-ORR, S., AYZENSHTAT, I., SHEFFER, M. & ALON, U. (2004) Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.
32. LOTITO, Q. F., MUSCIOTTO, F., MONTRESOR, A. & BATTISTON, F. (2022) Higher-order motif analysis in hypergraphs. *Commun. Phys.*, **5**, 79.
33. LOTITO, Q. F., MUSCIOTTO, F., BATTISTON, F. & MONTRESOR, A. (2022) Exact and sampling methods for mining higher-order motifs in large hypergraphs. *arXiv, arXiv:2209.10241*, preprint, not peer reviewed.
34. FORTUNATO, S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.
35. ZHOU, D., HUANG, J. & SCHÖLKOPF, B. (2006) Learning with hypergraphs: clustering, classification, and embedding. *In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)*. Cambridge, MA, USA: MIT Press, pp. 1601–1608.
36. CONTISCIANI, M., BATTISTON, F. & DE BACCO, C. (2022) Inference of hyperedges and overlapping communities in hypergraphs. *Nat. Commun.*, **13**, 7229.
37. RUGGERI, N., CONTISCIANI, M., BATTISTON, F. & DE BACCO, C. (2022) Generalized inference of mesoscale structures in higher-order networks.
38. LOTITO, Q. F., MUSCIOTTO, F., MONTRESOR, A. & BATTISTON, F. (2023) Hyperlink communities in higher-order networks. *arXiv, arXiv:2303.01385*, preprint, not peer reviewed.
39. TUDISCO, F. & HIGHAM, D. J. (2023) Core-periphery detection in hypergraphs. *SIAM J. Math. Data Sci.*, **5**, 1–21.
40. MICCICHÈ, S., MANTEGNA, R. N. ET AL. (2019) A primer on statistically validated networks. *Comput. Soc. Sci. Complex Syst.*, **203**, 91.
41. MUSCIOTTO, F., BATTISTON, F. & MANTEGNA, R. N. (2021) Detecting informative higher-order interactions in statistically validated hypergraphs. *Commun. Phys.*, **4**, 1–9.
42. MUSCIOTTO, F., BATTISTON, F. & MANTEGNA, R. N. (2022) Identifying maximal sets of significantly interacting nodes in higher-order networks. *arXiv, arXiv:2209.12712*, preprint not peer reviewed.
43. LANCICHINETTI, A., FORTUNATO, S. & RADICCHI, F. (2008) Benchmark graphs for testing community detection algorithms. *Physical Rev. E*, **78**, 046110.
44. NEWMAN, M. E. & GIRVAN, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.

45. CHODROW, P. S. (2020) Configuration models of random hypergraphs. *J. Complex Netw.*, **8**, cnaa018.
46. RUGGERI, N., BATTISTON, F. & DE BACCO, C. (2022) A principled, flexible and efficient framework for hypergraph benchmarking. *arXiv*, *arXiv:2212.08593*, preprint, not peer reviewed.
47. PETRI, G. & BARRAT, A. (2018) Simplicial activity driven model. *Phys. Rev. Lett.*, **121**, 228301.
48. BARRAT, A., BARTHELEMY, M. & VESPIGNANI, A. (2008) *Dynamical Processes on Complex Networks*. Cambridge University Press.
49. COUTINHO, B. C., WU, A.-K., ZHOU, H.-J. & LIU, Y.-Y. (2020) Covering problems and core percolations on hypergraphs. *Phys. Rev. Lett.*, **124**, 248301.
50. CARLETTI, T., BATTISTON, F., CENCETTI, G. & FANELLI, D. (2020) Random walks on hypergraphs. *Phys. Rev. E*, **101**, 022308.
51. SCHAUB, M. T., BENSON, A. R., HORN, P., LIPPNER, G. & JADBABAIE, A. (2020) Random walks on simplicial complexes and the normalized hodge 1-laplacian. *SIAM Rev.*, **62**, 353–391.
52. CARLETTI, T., FANELLI, D. & NICOLETTI, S. (2020) Dynamical systems on hypergraphs. *J. Phys.*, **1**, 035006.
53. MUOLO, R., GALLO, L., LATORA, V., FRASCA, M. & CARLETTI, T. (2023) Turing patterns in systems with high-order interactions. *Chaos, Solitons Fractals*, **166**, 112912.
54. GAMBUZZA, L. V., DI PATTI, F., GALLO, L., LEPRI, S., ROMANCE, M., CRIADO, R., FRASCA, M., LATORA, V. & BOCCALETTI, S. (2021) Stability of synchronization in simplicial complexes. *Nat. Commun.*, **12**, 1–13.
55. LUCAS, M., CENCETTI, G. & BATTISTON, F. (2020) Multiorder Laplacian for synchronization in higher-order networks. *Phys. Rev. Res.*, **2**, 033410.
56. MILLÁN, A. P., TORRES, J. J. & BIANCONI, G. (2020) Explosive higher-order Kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.*, **124**, 218301.
57. SKARDAL, P. S. & ARENAS, A. (2020) Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.*, **3**, 1–6.
58. ZHANG, Y., LUCAS, M. & BATTISTON, F. (2023) Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. *Nat. Commun.*, **14**, 1605.
59. DE ARRUDA, G. F., PETRI, G. & MORENO, Y. (2020) Social contagion models on hypergraphs. *Phys. Rev. Res.*, **2**, 023032.
60. IACOPINI, I., PETRI, G., BARRAT, A. & LATORA, V. (2019) Simplicial models of social contagion. *Nat. Commun.*, **10**, 1–9.
61. ST-ONGE, G., SUN, H., ALLARD, A., HÉBERT-DUFRESNE, L. & BIANCONI, G. (2021) Universal nonlinear infection kernel from heterogeneous exposure on higher-order networks. *Phys. Rev. Lett.*, **127**, 158301.
62. ALVAREZ-RODRIGUEZ, U., BATTISTON, F., DE ARRUDA, G. F., MORENO, Y., PERC, M. & LATORA, V. (2021) Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.*, **5**, 586–595.
63. CIVILINI, A., ANBARCI, N. & LATORA, V. (2021) Evolutionary game model of group choice dilemmas on hypergraphs. *Phys. Rev. Lett.*, **127**, 268301.
64. CIVILINI, A., SADEKAR, O., BATTISTON, F., GÓMEZ-GARDENES, J. & LATORA, V. (2023) Explosive cooperation in social dilemmas on higher-order networks. *arXiv*, *arXiv:2303.11475*, preprint, not peer reviewed.
65. GALLO, L., MUOLO, R., GAMBUZZA, L. V., LATORA, V., FRASCA, M. & CARLETTI, T. (2022) Synchronization induced by directed higher-order interactions. *Commun. Phys.*, **5**, 263.
66. GELARDI, V., GODARD, J., PALERESSOMPOULLE, D., CLAUDIERE, N. & BARRAT, A. (2020) Measuring social networks in primates: wearable sensors versus direct observations. *Proc. R. Soc. A*, **476**, 20190737.
67. G'ENOIS, M. & BARRAT, A. (2018) Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.*, **7**, 11.
68. GÉNOIS, M., VESTERGAARD, C. L., FOURNET, J., PANISSON, A., BONMARIN, I. & BARRAT, A. (2015) Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Netw. Sci.*, **3**, 326–347.
69. MASTRANDREA, R., FOURNET, J. & BARRAT, A. (2015) Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS One*, **10**, e0136497.

70. STEHLÉ, J., VOIRIN, N., BARRAT, A., CATTUTO, C., ISELLA, L., PINTON, J-F., QUAGGIOTTO, M., VAN DEN BROECK, W., RÉGIS, C., LINA, B. ET AL. (2011) High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One*, **6**, e23176.
71. VANHEMS, P., BARRAT, A., CATTUTO, C., PINTON, J-F., KHANAFER, N., RÉGIS, C., KIM, B.-A, COMTE, B. & VOIRIN, N. (2013) Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One*, **8**, e73970.
72. (2021) <https://journals.aps.org/datasets>.
73. BENSON, A. R., ABEBE, R., SCHAUB, M. T., JADBABAIE, A. & KLEINBERG, J. (2018) Simplicial closure and higher-order link prediction. *Proc. Nat. Acad. Sci.*, **115**, E11221–E11230.
74. SINHA, A., SHEN, Z., SONG, Y., MA, H., EIDE, D., HSU, B.-J. P. & WANG, K. (2015) An overview of microsoft academic service (MAS) and applications. *Proceedings of the 24th International Conference on World Wide Web*. ACM Press.
75. EPSTEIN, L., WALKER, T. G., HENDRICKSON, N. S. S. & ROBERTS, J. (2019) The U.S. Supreme Court Justices Database
76. LESKOVEC, J., KLEINBERG, J. & FALOUTSOS, C. (2007) Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, **1**, 2–es. <https://doi.org/10.1145/1217299.1217301>.
77. YIN, H., BENSON, A. R., LESKOVEC, J. & GLEICH, D. F. (2017) Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. New York, NY, USA: Association for Computing Machinery, pp. 555–564. <https://doi.org/10.1145/3097983.3098069>
78. BAUER-MEHREN, A., BUNDSCHUS, M., RAUTSCHKA, M., MAYER, M. A., SANZ, F. & FURLONG, L. I. (2011) Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One*, **6**, 1–13.
79. GALLO, L., LACASA, L., LATORA, V. & BATTISTON, F. (2023) Higher-order correlations reveal complex memory in temporal hypergraphs. *arXiv*, *arXiv:2303.09316*, preprint, not peer reviewed.
80. LUPPI, A. I., MEDIANO, P. A. M., ROSAS, F. E., HOLLAND, N., FRYER, T. D., O'BRIEN, J. T., ROWE, J. B., MENON, D. K., BOR, D. & STAMATAKIS, E. A. (2022) A synergistic core for human brain evolution and cognition. *Nat. Neurosci.*, **25**, 771–782.
81. GRILLI, J., BARABÁS, G., MICHALSKA-SMITH, M. J. & ALLESINA, S. (2017) Higher-order interactions stabilize dynamics in competitive network models. *Nature*, **548**, 210–213.