UNIVERSITÀ
DI TRENTO

The structural organization of higher-order networks

Measures, algorithms and software for hypergraphs

Quintino Francesco Lotito

Advisor: Prof. Alberto Montresor
Co-advisor: Prof. Federico Battiston
PhD Cycle: XXXVII

Abstract

Traditional network models have revealed important properties of real-world systems, including small-world and clustering organization. Yet, they often face limitations in capturing the full complexity of real-world systems. Over the years, researchers have devoted a lot of attention to expanding the set of frameworks available to model complex systems, incrementally encoding more features about the interactions, such as direction, time and multiplexity. Mounting empirical evidence, from metabolic interactions to scientific publishing, now suggests that real-world interactions tend to involve more than two units simultaneously. Explicitly encoding such *higher-order interactions* in our network modeling with hypergraphs or other tools seems a natural progression in network science research, will this be the key to new insights hidden by traditional binary models?

In this Thesis, we introduce computational and methodological tools to analyze the structural organization of systems with higher-order interactions at multiple scales. At the microscale, we propose efficient algorithms for higher-order motif analysis in hypergraphs, identifying overrepresented patterns of group interactions. At the mesoscale, we introduce hyperlink communities, which naturally capture both the hierarchical organization of hyperedges and community overlap in higher-order networks. To study group interactions with richer features, we introduce measures to analyze directed hypergraphs, including reciprocity definitions and motif analysis, providing insights into systems like metabolic networks and financial transactions. We also develop tools to study systems where nodes participate in different types of group interactions simultaneously, such as scientific collaborations across fields. To facilitate the adoption of higher-order network analysis, we develop Hypergraphx, an open-source Python library providing tools for hypergraph construction, visualization, and analysis. We complement this with Hypergraph-data, a curated repository of real-world datasets with rich metadata spanning different domains. These contributions aim to lower the entry barrier for higher-order network analysis and foster interdisciplinary collaboration.

With this Thesis, we aim to advance our ability to characterize and understand complex systems beyond traditional network approaches, opening new perspectives for modeling and analyzing group interactions in real-world systems.

Publications

This thesis builds upon the ideas, results and figures from the publications:

1. Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "Higher-order motif analysis in hypergraphs," *Communications Physics*, vol. 5, no. 1, p. 79, 2022
2. Q. F. Lotito, F. Musciotto, F. Battiston, and A. Montresor, "Exact and sampling methods for mining higher-order motifs in large hypergraphs," *Computing*, vol. 106, no. 2, pp. 475–494, 2024
3. 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*, vol. 11, no. 3, p. cnad019, 2023
4. Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "Hyperlink communities in higher-order networks," *Journal of Complex Networks*, vol. 12, no. 2, p. cnae013, 2024
5. Q. F. Lotito, A. Montresor, and F. Battiston, "Multiplex measures for higher-order networks," *Applied Network Science*, vol. 9, no. 1, p. 55, 2024
6. Q. F. Lotito, A. Vendramini, A. Montresor, and F. Battiston, "The microscale organization of directed hypergraphs," *arXiv preprint arXiv:2410.16258*, 2024
7. Q. F. Lotito, L. Betti, B. Nortier, A. Montresor, and F. Battiston, "Hypergraph-data: a repository for higher-order network data," *Working paper*, 2025

Other publications not covered in the thesis:

1. B. Arregui-García, A. Longa, Q. F. Lotito, S. Meloni, and G. Cencetti, "Patterns in temporal networks with higher-order egocentric structures," *Entropy*, vol. 26, no. 3, p. 256, 2024
2. S. Genetti, E. Ribaga, E. Cunegatti, Q. F. Lotito, and G. Iacca, "Influence maximization in hypergraphs using multi-objective evolutionary algorithms," in *International Conference on Parallel Problem Solving from Nature*, pp. 217–235, Springer, 2024

Table of contents

1 | Introduction

The study of complex systems requires sophisticated mathematical and computational tools to understand how collective behaviors emerge from the shape of the interactions among system components. This chapter introduces the fundamental concepts and frameworks that form the foundation of this thesis. It highlights both the strengths and limitations of traditional network approaches, motivates and formalizes higher-order network representations and provides an overview of the thesis's structure and main contributions.

1.1 The backbone of complex systems

One of the most distinguishing features of *complex systems* is the intricate patterns of interactions among their units. Rather than existing as isolated entities, the fundamental components of such systems, including individuals in a social group, neurons in a brain, species in an ecosystem, or web pages on the internet, are deeply and dynamically interconnected. Such connections give rise to behaviors, properties and dependencies that cannot be fully understood by examining individual units in isolation.

The conceptual and mathematical framework of network theory provides a powerful lens through which these connectivity patterns can be examined, understood, and ultimately leveraged to gain insight into the emergent phenomena of complex systems. In its most general sense, a *network* consists of nodes representing the system's primary elements and edges that characterize the interactions, relationships or dependencies between these elements. Networks serve as abstract representations that can transcend disciplinary boundaries: the same tools and concepts can be applied to uncover the universal properties of seemingly disparate systems. For instance, the way neurons form synaptic connections in the human brain resembles how people forge social ties in a community, how genes regulate each other's expression in a cell, or how financial institutions trade assets and liabilities. In all these settings, networks provide a unifying language that captures both the complexity of the local interactions and the broader, often counterintuitive, global outcomes that arise [10, 11].

This dual perspective, where micro-level structure shapes macro-level outcomes, reveals the compelling nature of networks. The topology of a network can strongly influence how quickly information spreads, how robust the system is to failures or attacks, and how new patterns of organization emerge. A tightly knit cluster of nodes might form a cohesive *community* with dense internal connections, while sparser links between communities can regulate the flow of resources, information, or influence [12]. Likewise, certain nodes, often referred to as *hubs*, may assume disproportionately large roles in shaping global dynamics [13], serving as bottlenecks or bridges in the flow of signals. By mapping such architectures, we can begin to explain why certain networks are remarkably resilient to random disruptions but vulnerable to targeted removals, or why some systems undergo sudden transitions from stable to unstable regimes. In Fig. 1.1, we use a network to model social encounters of 327 students from 9 classes in a High School, where nodes represent the individual students and edges correspond to their face-to-face interactions across multiple days [14]. We highlight the emergence of communities that tend to align with the classes of the students, suggesting that students primarily interact within their own class groups and reflecting the natural clustering of social ties based on shared environments.

From a historical perspective, the formal study of networks gained momentum as empirical data became more accessible and computational tools more powerful. Early theoretical work on *random graphs* [16, 17] by Erdős and Rényi laid a foundational understanding of how large-scale connectivity emerges from simple probabilistic rules. Later research introduced more structured and realistic network models, such as *small-world* networks [18] demonstrating high clustering and short path lengths and *scale-free* [13] networks emphasizing the importance of hubs. Today, a rich toolkit exists, encompassing graph theory, statistical mechanics, dynamical systems, and data science, enabling researchers to characterize network structure, model dynamic processes on networks, and identify principles that apply across domains.

The study of networks thus becomes a cornerstone in the broader effort to decode the complexity of real-world systems. By employing network-based methods, we seek not only to describe patterns of connectivity but also to understand how those patterns influence the collective behavior of the system. This understanding is crucial when we aim to manipulate system properties such as controlling epidemic outbreaks, optimizing the resilience of infrastructure, improving communication in distributed teams, or fostering innovation in economic networks. Ultimately, networks are not merely convenient metaphors; they are the structural backbone upon which complexity rests.

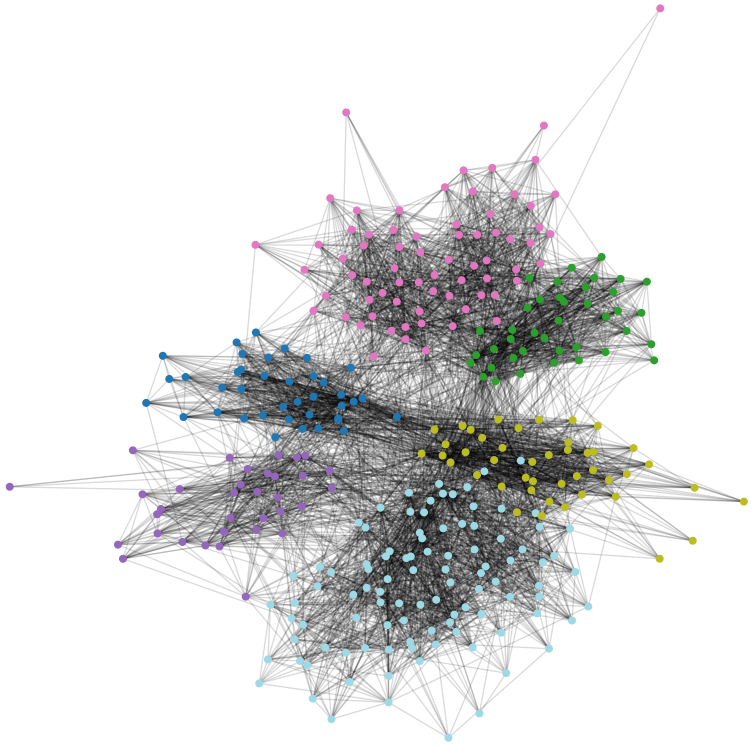

Figure 1.1: Network of students' interactions in a high school across multiple days. Nodes represent students, edges represent face-to-face encounters, and node colors indicate distinct communities identified using a modularity maximization algorithm [15].

1.2 Limitations of traditional network models

Having established the fundamental idea of representing complex systems as networks of interconnected entities, we must now acknowledge that the initial models we considered, which consisted of static, undirected, unweighted binary edges, are only the simplest manifestations of a far richer conceptual framework. While these traditional network representations have significantly advanced our understanding of collective phenomena, from social contagion to the robustness of infrastructure, they often remain too coarse to capture the full complexity of real-world systems.

One key limitation lies in the assumption that all links are symmetrical and unchanging. In many domains, interactions are inherently *directed* [19], as with citation networks or communication flows, where information or influence tends to travel along preferred pathways rather than moving freely in both directions. Similarly, most systems do not remain static over time. Instead, connections appear, disappear, or change in strength, giving rise to *temporal* networks that evolve in response to external factors or internal dynamics [20]. Ignoring this temporal dimension can obscure early-warning signals, critical transient states and patterns that only emerge when we view connectivity as an ongoing process rather than a fixed snapshot [20].

Additionally, real systems often involve multiple types of relationships that overlap and interact [21]. A single pair of entities might be simultaneously connected by friendship, economic exchange, and shared membership in an organization. Traditional models, which typically consider only one type of link at a time, fail to reflect the interplay of these parallel connections. *Multilayer* (or multiplex) networks address this shortcoming by integrating several distinct interaction layers into a unified representation. This approach reveals how different dimensions reinforce or counteract one another, producing behaviors that single-layer analyses cannot anticipate [22].

In Fig. 1.2a, we present an example of a system consisting of two entities exchanging messages across multiple communication channels. Modeling this system as a plain graph results in a significant loss of information (Fig. 1.2b), such as the number of messages exchanged, the source and destination of each message, their temporal sequence, and the specific communication channel used. To accurately capture the complexity of the system, it is essential to use a more sophisticated mathematical model capable of encoding all these interaction features, preserving the rich structure and dynamics of the observed system (Fig. 1.2c).

At this point, we have introduced a wide array of tools to analyze real-world systems.

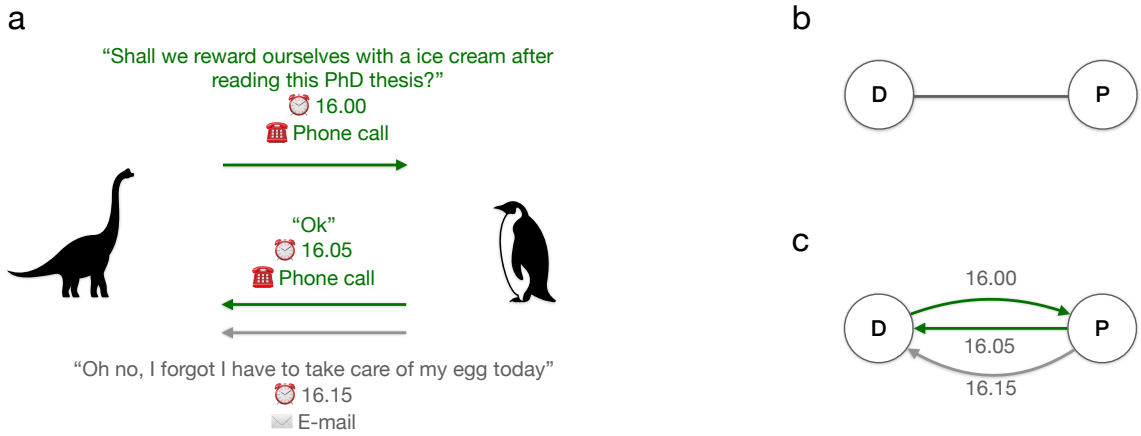

Figure 1.2: Illustration of a system where two entities exchange messages across multiple communication channels (a). Modeling the system as a plain graph (b) leads to a significant loss of information, such as the number, direction, temporal order, and communication channel of the messages. A more sophisticated mathematical model (c) is required to preserve the rich structure and dynamics of the system.

These tools extend beyond simple networks, introducing a tricky trade-off between the complexity of the model and its ability to capture the essential aspects of the systems under study. It is therefore worth stressing that encoding every detail of the available information is not always necessary or practical. The true strength of such advanced network tools lies in their flexibility, allowing researchers and practitioners to tailor the level of complexity of the model to suit the specific needs of their analysis, ensuring that critical information is preserved without overburdening the model.

1.3 Higher-order interactions in the real-world

Even as we incorporate direction, time, and multiple layers of connectivity, we remain within the paradigm of pairwise relations. That is, each edge still connects precisely two nodes. While these enriched frameworks bring us closer to reality, certain systems defy such pairwise constraints. In fact, interactions in complex systems often occur among groups of more than two entities at once [23, 24]. The next step in this conceptual journey is to move beyond networks and consider mathematical representations allowing *higher-order* connectivity.

In recent years, the interest in investigating the significance of higher-order interactions in real-world systems has steadily grown. From social and biological to technological

domains, evidence suggests that many collective phenomena emerge from simultaneous interactions among multiple system units rather than just pairs. In social systems, for instance, face-to-face interactions frequently occur in groups of varying sizes, from small conversations to larger gatherings [25]. Similarly, in scientific collaboration networks, research papers often result from the joint effort of multiple authors working together as a cohesive unit, rather than through a collection of independent binary collaborations [26]. In biological systems, higher-order interactions are fundamental to understanding system behavior. Cellular networks exhibit complex metabolic reactions where multiple genes or proteins interact simultaneously to produce specific outcomes [27]. Brain networks demonstrate coordinated firing patterns among groups of neurons that cannot be decomposed into simple pairwise relations [28]. In ecological systems, species interactions often involve multiple participants, such as in food webs where predator-prey relationships can involve multiple species competing for resources or collaborating in hunting [29]. Technological and communication systems also manifest higher-order patterns. Email communications frequently involve multiple recipients, creating group-level information exchange patterns [30]. In financial networks, transactions can involve multiple parties simultaneously, as seen in cryptocurrency systems where multiple wallets may participate in a single transaction [31].

These examples underscore how higher-order interactions are not mere abstractions but fundamental features of real-world systems that shape their function and dynamics.

1.4 Representing higher-order interactions

Traditional network models, while powerful in their simplicity, face fundamental limitations when representing group interactions [23, 24]. In these models, edges can only connect pairs of nodes, forcing any group interaction to be artificially decomposed into a collection of pairwise relationships. Consider, for instance, a scientific paper authored by three researchers. A traditional network would represent this collaboration as three separate edges forming a triangle, suggesting three independent pairwise collaborations rather than a single cohesive group effort. This representation not only fails to capture the unified nature of the interaction but can also lead to misleading analyses. When each group interaction is expanded into a clique (*clique-projection*), the network becomes artificially dense with edges that do not represent real independent relationships. This distortion can significantly impact various network measures, from local clustering coefficients to community structure, potentially leading to incorrect interpretations of the system's organization and dynamics. To overcome these limitations, researchers have

turned to higher-order network representations, such as *hypergraphs* and *simplicial complexes*. These mathematical frameworks allow interactions to be modeled explicitly as group-level entities, rather than being decomposed into pairwise edges.

In a hypergraph, nodes can be connected by hyperedges, where each hyperedge can involve any number of nodes simultaneously. For example, a research paper authored by three individuals can be represented as a single hyperedge connecting all three nodes, preserving the integrity of the group interaction. Simplicial complexes offer another powerful tool for capturing higher-order relationships. In these structures, interactions are represented as simplices: a single node is a 0-simplex, a pairwise edge is a 1-simplex, and a group of three nodes forms a 2-simplex (a triangular face), and so on. Importantly, in simplicial complexes higher-order interactions build upon lower-order ones in a consistent and mathematically grounded way, in fact, they impose a hierarchical organization on interactions, where every higher-order interaction inherently includes all lower-order interactions among its participating nodes. For example, a 2-simplex (triangle) implies the existence of three 1-simplices (edges) between each pair of nodes.

While both hypergraphs and simplicial complexes capture higher-order interactions, they differ in their representation and interpretation of these interactions. Hypergraphs are more flexible in defining group interactions, as hyperedges can connect any arbitrary subset of nodes without enforcing hierarchical constraints. This flexibility makes hypergraphs well-suited for representing systems with diverse and irregular group interactions, such as collaborative projects or multi-agent communications. However, this generality comes at the cost of reduced mathematical structure, limiting the availability of certain analytical tools. In contrast, the hierarchical nature of simplicial complexes enables using powerful tools from algebraic topology (e.g., persistent homology) to analyze higher-order structures but at the cost of less flexibility for representing arbitrary group interactions.

Bipartite graphs are another popular framework for encoding higher-order interactions. They consist of two distinct sets of nodes, with edges connecting only nodes from different sets. One set typically represents system units, while the other represents interactions within the system. An edge between two nodes indicates the participation of a unit in a group interaction. Bipartite graphs are valuable for their simplicity and resemblance to traditional graphs, and for historical reasons, they are particularly popular in fields like ecology. Additionally, they allow for straightforwardly encoding features on edges, for instance, temporal information to track a unit's entry and exit from a group interaction. This makes bipartite graphs a useful data structure for representing higher-order networks in software and simplifying algorithm design. However, the need to define two types of nodes, i.e., system units and polyadic interactions, can complicate the expression of

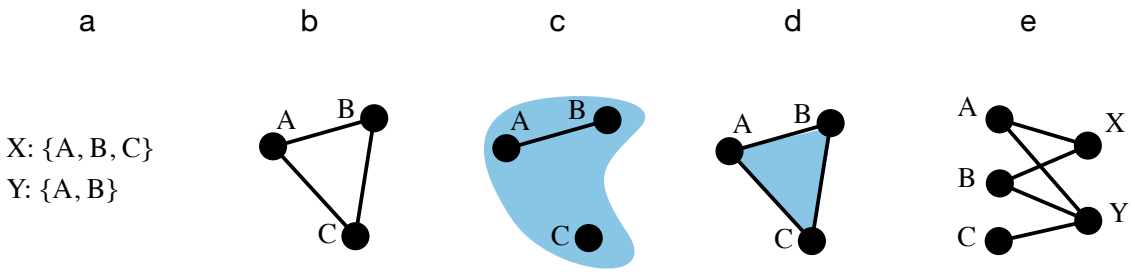

Figure 1.3: A toy example of a system with higher-order interactions, including a three-way interaction, represented across various frameworks: (b) a clique-projection, where group information is lost as the triangle can be interpreted as either a set of binary interactions or a three-way interaction; (c) a hypergraph which represents interactions without loss of information; (d) a simplicial complex which encodes both a three-way interaction but also introduce binary interactions that are not observed in the original system; (e) a bipartite graph which encodes the system without losing information but at the cost of introducing nodes with two different meaning.

traditional network science problems, which focus on direct unit-to-unit interactions. This is why hypergraphs have gained more attention in network science. Hypergraphs, in certain cases, offer a more natural and efficient framework for algorithmic solutions. For example, in Chapter 2, we explore motifs in higher-order networks, where hypergraphs provide a natural way to define and interpret recurring patterns of group interactions among a fixed number k of units. In contrast, in bipartite graphs the number of nodes is not clearly defined and not fixed, leading to a more complicated analysis and interpretation of the results. Moreover, a hypergraph formalization of the problem naturally leads to a very efficient mining algorithm.

In Fig. 1.3, we present a toy example of a system with higher-order interactions, including a three-way interaction, and demonstrate how the system can be represented using different frameworks. The figure compares the representations of the system through clique-projections, hypergraphs, simplicial complexes, and bipartite graphs, highlighting the trade-offs between information preservation and the complexity of node interpretations in each framework. In particular, we show that the clique-projection cannot retain information about group interactions, because for example the triangle can be interpreted either as a set of binary interactions or as a three-way interaction. The simplicial complex captures the three-way interaction but also introduces binary interactions that are not observed in the original system. The hypergraph and the bipartite graph preserve all information from the system, but the bipartite graph introduces nodes with dual meanings.

	Description	Pros	Cons	Best for
Hypergraphs	Nodes are connected by hyperedges, each involving multiple nodes	Captures arbitrary group interactions; flexible representation	Lacks hierarchical constraints, limiting some mathematical tools	Analyzing multi-agent interactions, group-based structures, network motifs
Simplicial Complexes	Interactions are defined hierarchically as simplices	Strong mathematical structure; useful in algebraic topology	Enforces hierarchical relationships, restricting arbitrary groups	Topological data analysis, modeling networks with clear hierarchical interactions
Bipartite Graphs	Two-node types (units and interactions) with edges between them	Preserves group information; allows structured encoding (e.g., temporal data)	Requires an extra node type, complicating interpretation of traditional graph measures and algorithms	Encoding higher-order interactions in a structured way, established in certain communities (e.g., ecological networks)
Clique-Projection	Group interactions are projected into pairwise edges	Compatible with traditional graph methods	Artificially increases density; loses group interaction information	When only pairwise relationships are needed, basic network analysis

Table 1.1: Comparison of higher-order network representations. Each framework has unique advantages and limitations, making them suitable for different applications.

All in all, there is not a clear "winner" among the available mathematical frameworks for encoding higher-order structures. Hypergraphs, simplicial complexes, and bipartite graphs each have their own strengths and limitations, making them more or less suitable depending on the specific application. Finally, it is important to observe that while certain problems may be more naturally described in one framework over another, also algorithmic solutions may favor a specific framework. For a more summarized and structured comparison of the different mathematical representations of higher-order interactions, we refer to Table 1.1.

In the following of this Thesis, we will mostly focus on hypergraphs.

1.5 Beyond the buzzword: when is "higher-order" useful?

The adoption of higher-order frameworks for modeling complex systems comes at a considerable cost, as it introduces significant computational challenges and requires the development of entirely new analytical tools. Moreover, the added complexity in data collection, storage, and processing raises legitimate questions: is this additional complexity justified by meaningful insights that simpler models fail to capture?

Many-body interactions, as we have seen in the previous sections, are quite common in nature. Therefore, we cannot run away from the fact that many real-world datasets are naturally hypergraphs. Traditional approaches consider lower-order projections of such datasets, where hyperedges are replaced by cliques, in order to get a network and apply classic results in network science. Most of the works on higher-order networks build on the intuition that this clique-expansion modeling can fail very brutally in some, not so rare, circumstances. Consider for instance Fig. 1.4. If we slightly modify the hypergraph on the left by adding a new large hyperedge, its clique-projection (on the right) gets heavily affected, as a large clique is introduced in the system. From a modeling perspective, adding this large clique destroys any structural organization and introduces artificial edges that are not independent of each other. From a computational perspective, it adds a huge number of new edges to consider in the computation. Higher-order approaches, by contrast, do not suffer from such drawbacks.

Building on this potential modeling advance, recent works in the field have begun to address the challenges that the higher-order frameworks bring by developing new mathematical and computational tools. First, fundamental network science concepts have been generalized to accommodate group interactions. These generalizations include central-

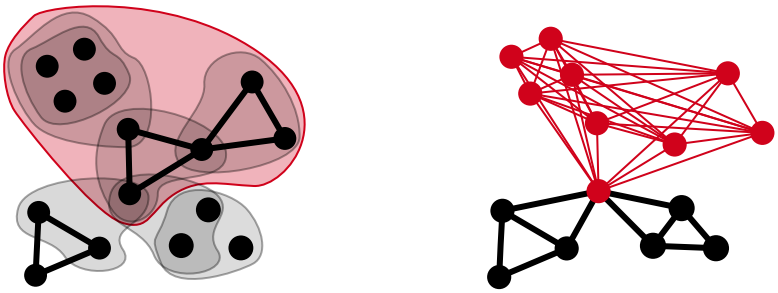

Figure 1.4: **Drawbacks of lower-order approaches to higher-order data.** If a large hyperedge (the red one) is added to a hypergraph, it can significantly affect its clique-projection, destroying the ability of low-order community detection tools to capture any structural organization and introducing a large number of new edges to consider in computation. These drawbacks are resolved when higher-order data are handled using a higher-order approach.

ity measures [32, 33], directedness [34], clustering [30, 35] and assortativity [36], just to mention a few. An explicit treatment of many-body interactions, including their inference and reconstruction [37], is necessary to understand network formation mechanisms [38–41], fully capture the real community structure of higher-order systems [42–44] and extract their statistically validated higher-order backbone [45]. Taking into account higher-order interactions revealed to be crucial to understand the emergent behavior of complex systems, as they have been found to profoundly impact diffusion [46, 47], synchronization [48–52], social [53–55] and evolutionary [56] processes. Higher-order network analysis allowed researchers to extract new insights into the properties and organization of several real-world complex systems characterized by a prevalence of many-body interactions, including collaboration networks [26], human face-to-face interactions [25], folksonomies [57], species interactions within complex ecosystems [29], brain networks [28], and cognitive associations [58].

Finally, in order to answer the question of this section, higher-order tools are useful when they produce novel insights or faster algorithms that cannot be obtained by considering lower-order projections of the data. It is important to acknowledge that not every system is interesting in this sense, as there are examples for which lower-order representations can reasonably approximate the more general higher-order ones [59]. However, in general, literature about the practical implications of higher-order modeling is steadily growing, with higher-order frameworks establishing themselves as important tools in the toolbox of a modern network scientist and new results in this area are contributing to the broader goal of understanding how collective behaviors emerge from complex patterns of connectivity.

1.6 Outline of the thesis

In this Thesis, we aim to contribute to expanding our current knowledge of systems with higher-order interactions, investigating their structural properties at multiple scales, with a particular focus on developing new computational tools and analytical frameworks for hypergraphs.

The rest of this Thesis is organized as follows:

Chapter 2 introduces higher-order motifs, extending the traditional concept of network motifs to hypergraphs. We develop efficient algorithms to extract recurrent patterns of higher-order interactions and demonstrate their utility in characterizing the local structure of real-world hypergraphs across different domains. Additionally, we introduce a sampling method to perform motif analysis on large-scale hypergraphs, balancing accuracy and running time.

Chapter 3 presents methods for analyzing hyperlink communities in hypergraphs, providing insights into the mesoscale organization of higher-order networks. This approach reveals how group interactions are hierarchically organized and how individual nodes participate in multiple communities simultaneously.

Chapter 4 explores properties of directed hypergraphs, introducing new measures to characterize their microscale organization. We develop tools to analyze patterns of directed higher-order interactions and investigate reciprocity in group-level relationships.

Chapter 5 extends our analysis to multiplex hypergraphs, introducing measures to characterize systems where multiple types of higher-order interactions coexist. We demonstrate how these tools reveal complex patterns of connectivity across different interaction layers.

Chapter 6 presents two contributions to expand the software and data ecosystem higher-order network research. First, we introduce Hypergraphx (HGX), a comprehensive Python library which implements the methods developed throughout this thesis alongside other tools from the literature. Second, we present hypergraph-data, a curated repository of datasets representing real-world systems with higher-order interactions from different domains. This repository addresses the growing need for accessible and standardized higher-order network data, facilitating reproducible research and benchmarking of new methods.

Finally, Chapter 7 summarizes our contributions and discusses future directions for research in higher-order network science, including the development of more sophisticated analytical tools and the application of our methods to emerging domains.

2 | Higher-order motif analysis in hypergraphs

Motifs are small, recurring subgraphs that appear in a network with a frequency significantly higher than expected in a random graph model [60]. These motifs are considered as the building blocks of network structure, providing valuable insights into the functional organization of complex systems. Over the years, the study of motifs has revealed characteristic patterns of interactions crucial to understanding the behavior of many networks across different domains [61], including the presence of functional circuits in gene regulatory networks [62], communication patterns in social networks [63, 64], and structural patterns in neural networks [65].

In this Chapter, we extend the traditional concept of network motifs to hypergraphs. We introduce efficient algorithms for extracting recurrent patterns of higher-order interactions and demonstrate their effectiveness in characterizing the local structure of real-world hypergraphs across various domains, including biological and social networks. Furthermore, we propose a novel sampling method designed to perform motif analysis on large-scale hypergraphs, striking a balance between computational efficiency and the accuracy of the extracted motifs.

2.1 The building blocks of complex networks

Networked systems may be differentiated by their preferential patterns of connectivity at the microscale, encoding a characteristic fingerprint often relevant for system functions. This may be quantified by measuring network motifs, small connected subgraphs that appear in an observed network at a frequency that is significantly higher than in a random-graph null model [60]. The analysis of the motifs of a network revealed the emergence of "superfamilies" of networks, i.e., clusters of networks which display similar local structure (see Fig. 2.1). These clusters tend to group networks from similar domains or networks that have evolved via similar evolutionary processes [61]. Motifs can be interpreted as elementary computational circuits, with specific functionalities that can be shared by

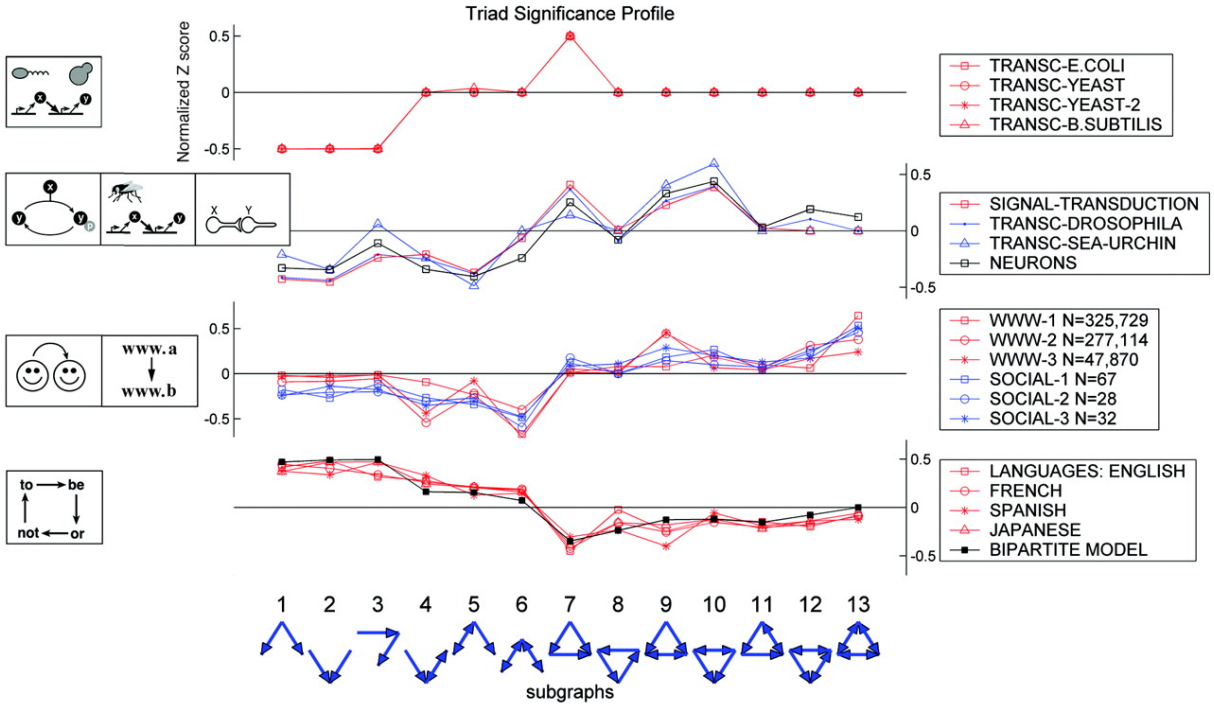

Figure 2.1: The analysis of motifs in real-world networks highlights the emergence of distinct patterns shared by networks from similar domains, enabling the identification of families of networks with similar local structures. Motifs are therefore interpreted as elementary computational circuits encoding specific functionalities. Each plot groups networks representing systems within the same domain. The z-score on the y-axis of each plot represents the abundance of each motif relative to a null model, with correlations observed between systems within the same domain, highlighting shared patterns and structural similarities. From [61]. Reprinted with permission from AAAS.

similar networks. For example, transportation networks are designed to simplify the traffic flow, whereas gene regulation and neuron networks are often thought to have evolved to process information. These functional differences in such systems are reflected in the emergence of different significant motifs in the networks that describe them. In this regard, studying motifs can also give new insights into the dynamics and resilience of classes of networks [61, 66]. To explicitly uncover the relation between the dynamical processes that unfold on a network and its structural decomposition at the local scale, recently a refined notion of process motifs has been proposed [67], introducing a framework to assess the contribution of each motif to the overall dynamical behavior of the system.

Network motifs have been used in a wide range of applications. In biology, motifs have been extensively studied for the analysis of transcription regulation networks (i.e. networks that control gene expression). Studies show that diverse organisms from bacteria

to humans exhibit common regulation patterns, each with its very own function in determining gene expression [62, 68–71]. Similarly, motif analysis has been applied to show how complex and flexible neural functions emerge from the composition of fundamental circuits in brain networks [65]. Moreover, motifs have also been used as a feature for the identification of cancer [72]. Eventually, the need to analyze biological datasets of ever-increasing size has been a strong motivation for the development of more efficient algorithms [73]. Beside biology, motifs have also been applied to provide fingerprints of the local structures of social networks [63, 64], for the early detection of crisis-leading structural changes in financial networks [74] and to study the networks of direct and indirect interactions across species in ecology [75, 76].

The interest of the research community in extracting fingerprints at the network microscale of real-world systems has led to considering richer frameworks for motif analysis [77], including extensions to more general network models such as weighted [78], temporal [20] and multilayer [22] networks. Weighted networks can be characterized in terms of the intensity and coherence of the link weights of their subgraphs [79]. Temporal networks can be studied at both topological and temporal micro- and mesoscale by considering time-restricted patterns of interactions [80, 81]. Statistically over-expressed small multilayer subgraphs [82] highlight the local structure of multilayer networks such as the human brain [83]. Nevertheless, the methods, algorithms and tools proposed in literature so far mostly consider only patterns of pairwise interactions, thus limiting our capabilities of characterizing the local structure of systems that involve group interactions. Recently, Lee et al. [84] made a first contribution to close this gap: at difference with traditional motif analysis that focuses on patterns of interactions among small sets of nodes, they investigated patterns associated with connected hyperedges, in particular the 26 possible ways in which 3 connected hyperedges can overlap, allowing to extract information on the design principles of hypergraphs. Instead, in this Chapter, we provide a general and scalable methodology which naturally generalize to hypergraphs the seminal notion and analysis of network motifs proposed by Milo et al. [60] for traditional graphs.

2.2 Foundations of motifs in hypergraphs

In order to systematically study the local structure of higher-order networks, we define higher-order motifs straightforwardly by extending the traditional definition from Milo et al. [60]. Higher-order motifs are patterns of small connected sub-hypergraphs that occur in an observed hypergraph at a frequency that is statistically significantly higher with respect to a null model. We refer to the number of nodes involved in the pattern

as the *order* of a higher-order motif. We will often use the word "motifs" to refer to all the possible patterns of sub-hypergraphs involving a certain number of nodes rather than only those over-expressed.

To perform a (higher-order) motif analysis of a system, we need to (i) count the frequency of each query (higher-order) motif in a graph/hypergraph, (ii) compare the frequency of each query (higher-order) motif with that observed in a null model, and (iii) evaluate the over- or under-expression of each higher-order motif.

In this Chapter, we are interested in different aspects related to motif analysis in hypergraphs. On a computational side, we study the combinatorial properties of the patterns of subhypergraphs and devise efficient algorithms for motif discovery of sizes 3, 4 and 5. In particular, extracting motifs of size 5 requires the careful design of a new sampling algorithm. We provide a performance evaluation of the algorithms applied to real-world datasets and an assessment of the accuracy of the approximated algorithm. From a network science perspective, we apply such algorithms to characterize real-world hypergraphs at their local scale and highlight the emergence of families of higher-order networks in a way similar to Milo et al. [61].

2.3 Combinatorics of higher-order motifs

The number of possible patterns of pairwise undirected interactions involving three connected nodes is only two, however, it grows to six when considering also higher-order interactions (Fig. 2.2a).

Finding an analytical form encoding the dependence of the number of higher-order motifs on the motif order k is a challenging task due to the constraints related to the computation of all possible combinations of higher-order interactions among k nodes. However, we are able to compute a upper and lower bounds for this number. We denote with m the number of all the possible non-isomorphic connected hypergraphs of k vertices (we recall that two hypergraphs are isomorphic if they are identical modulo relabeling of the vertices). To compute an upper bound on m, we can count the number of labelled hypergraphs ignoring the constraint on being non-isomorphic and connected. There are $\binom{k}{i}$ possible hyperedges of size i over k vertices. We are interested only in the hyperedges with cardinality at least 2, therefore there are $\sum_{i=2}^{k}\binom{k}{i} = 2^k - k - 1$ possible hyperedges. When creating a labelled hypergraph we can either include each hyperedge or not, this yields a total number of possible labelled hypergraphs equal to 2^{2^k-k-1}. To compute the lower bound of m, we construct connected hypergraphs on k vertices as follows. First, we pick any chain of edges and put all the edges in the hypergraph. This uses $k-1$ edges and makes sure

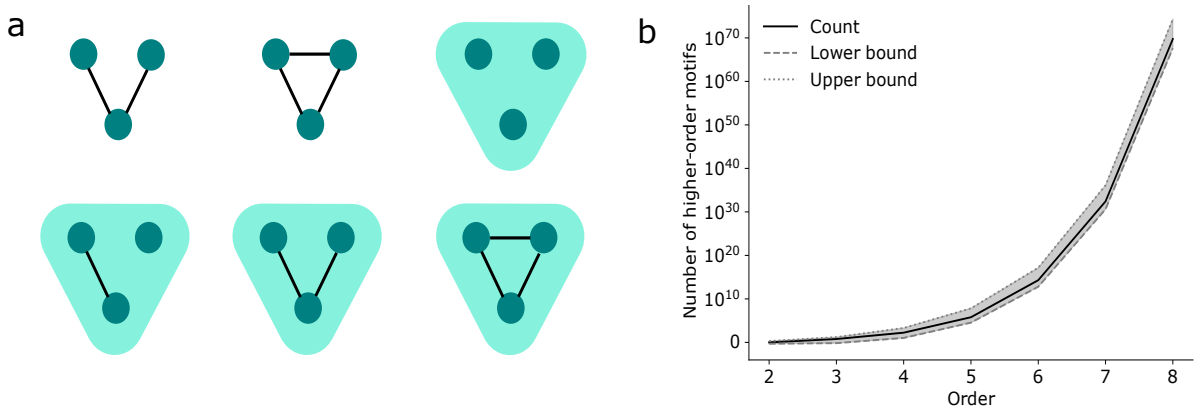

Figure 2.2: **Combinatorics of higher-order motifs. a)** Enumeration of all the six possible patterns of higher-order interactions involving three nodes. Green shaded triangles represent higher-order interactions, whereas black lines represent pairwise interactions. **b)** Upper and lower bounds on the number of higher-order motifs as a function of the order (gray shaded area). The black line represents the exact count for small orders.

the hypergraph is connected. There are $(2^k - k - 1) - (k - 1) = 2^k - 2k$ potential edges left over. Each of those edges can be added or not to the hypergraph, yielding at least 2^{2^k-2k} connected hypergraphs. However, we have to count only non-isomorphic copies, and have so far counted labelled graphs. For each unlabelled graph, there are at most $k!$ ways of labeling the vertices. So the number of non-isomorphic connected hypergraph is at least $\frac{2^{2^k-2k}}{k!}$. Fig. 2.2b shows the upper and lower bounds on the growth of the possible higher-order motifs as a function of the order, as well as the exact count for small orders, showing that this function has a super-exponential growth.

2.4 Mining higher-order motifs in hypergraphs

The enumeration of all the patterns of connected sub-hypergraphs of a given size is obviously the most expensive sub-task in motif analysis. The weight of this step is even more impactful considering that it must be repeated in randomized networks as well. To solve this problem exactly, in this section, we propose a baseline algorithm based on projecting the hypergraph onto a graph and employing state-of-the-art motif analysis algorithms on it. Additionally, we present a more efficient method that directly leverages higher-order structures to construct sub-hypergraphs of a specified size.

2.4.1 Basic definitions

Before diving into the algorithmic aspects of the problem of motif discovery, we introduce some tools that will be useful later on.

Definition 1 (Vertex-induced sub-hypergraph)**.** *The sub-hypergraph $\mathcal{H}[V']$ induced by the subset $V' \subseteq V$ is the pair (V', E'), where $E' = \{e \in E : e \subseteq V'\}$.*

Now we can define more formally the notions related to the *isomorphism* problem, which is the fundamental theoretical tool underlying network motif discovery.

Definition 2 (Hypergraph isomorphism)**.** *Two hypergraphs $\mathcal{H} = (V, E)$ and $\mathcal{H}' = (V', E')$ are isomorphic if they are identical modulo relabeling of the vertices. More formally, if there exists a bijection $f : V \to V'$ such that $e = \{u_1, ..., u_n\} \in E$ if and only if $e' = \{f(u_1), ..., f(u_n)\} \in E'$.*

This allows us to define the notion of occurrence, which is central to the problem of motif discovery.

Definition 3 (Occurrence)**.** *Given a hypergraph $\mathcal{H} = (V, E)$ and a smaller query hypergraph $\mathcal{Q} = (V', E')$, the occurrences of $\mathcal{Q}$ in $\mathcal{H}$ are all the sub-hypergraphs of $\mathcal{H}$ isomorphic to $\mathcal{Q}$. We often refer to the number of occurrences of $\mathcal{Q}$ in $\mathcal{H}$ as the* frequency *of $\mathcal{Q}$ in $\mathcal{H}$.*

All in all, the problem of motif discovery involves finding and counting the occurrences of all the possible patterns of subhypergraphs given a certain number of nodes.

2.4.2 Baseline

While traditional algorithms are not able to identify patterns of polyadic interactions, they can be used as a routine for more sophisticated algorithms. In our baseline, we consider the projected graph of a hypergraph. We recall that the projection of a hypergraph $\mathcal{H} = (V, E)$ is a graph $G = (V, E')$, defined on the same vertices of $\mathcal{H}$ and such that an edge between two vertices $a, b \in V$ exists if and only if a and b participate together in at least a hyperedge $e \in E$. In other words, every hyperedge $e \in E$ is replaced in G with a clique.

By running a classic algorithm (e.g., ESU [85]), we can efficiently enumerate connected subgraphs of size k in the projected graph. However, these subgraphs are only *candidate higher-order motifs* for two potential reasons: (i) they do not include higher-order interactions; (ii) even if a subgraph s of size k is connected in the projected graph, the sub-hypergraph induced by the vertices in s and the hyperedges E may be not connected.

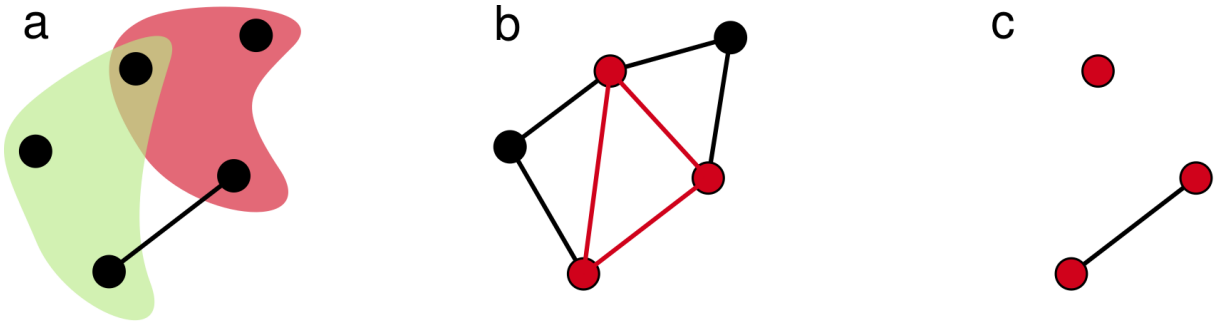

Figure 2.3: a) Example of a hypergraph H in which the baseline fails. b) We highlight in red a connected subgraph s of size $k = 3$, one of the many possible outputs of a standard motif discovery algorithm applied on the projection of the previous hypergraph. c) The sub-hypergraph induced by the vertices of s and the hyperedges of H is not connected.

We highlight these pitfalls in Figure 2.3.

In order to account for these issues, we construct the sub-hypergraph induced by the k nodes of the candidate motif (see Section 2.4.4) and check if this sub-hypergraph is connected. If it is, then one can simply update the frequency hash map, otherwise, the output is discarded. A more formal explanation of this method is reported in Algorithm 2.1. All in all, the baseline inherits the complexity of ESU [85], which in the worst case is $O(N^k)$, where N is the number of nodes in the network and k is the size of the subgraphs. However, in practical graphs (e.g., social and biological networks), the actual number of connected subgraphs is typically much smaller, making ESU feasible for moderate values of k. Additionally, there is the preprocessing cost of computing the clique projection of the hypergraph, which is $O(|E| \cdot k^2)$, where $|E|$ is the number of hyperedges and k denotes the average hyperedge size.

Algorithm 2.1 Baseline: Counting higher-order motifs

Input: a hypergraph $\mathcal{H} = (V, E)$ and an integer $k \in \{3, 4\}$.
Output: distribution of the frequency of the motifs of order k.

Let $G = (V, E')$ be the projection of $\mathcal{H}$
Let $\mathcal{M}$ be the motifs frequency dictionary
$\mathcal{S} \leftarrow \text{ESU}(G, k)$
for each *subgraph* $= (V^*, E^*) \in \mathcal{S}$ **do**
    candidate_motif $\leftarrow$ sub-hypergraph (of $\mathcal{H}$) induced by V^*
    if *candidate_motif* is connected **then**
        Let $\mathcal{C}_m$ be the isomorphism class of *candidate_motif*
        $\mathcal{M}[\mathcal{C}_m]$ += 1
    end if
end for

2.4.3 Efficient algorithms

The most expensive step in the previous algorithm is obviously the ESU subroutine. Moreover, the performance is widely impacted by the fact that hypergraph projections can be very dense and that lots of subgraphs are discarded for not satisfying the requirement of the induced sub-hypergraphs of being connected. To solve these problems, we work directly on hypergraphs, designing an efficient algorithm that exploits containment properties of higher-order structures in real-world systems. We optimize separately the two cases of 3- and 4-node motifs.

3-node motifs As shown in Figure 2.2, two of the motifs involving three nodes are composed only by pairwise relations, while the others involve one hyperedge of order 3. To discover the latter, it is enough to iterate over all the hyperedges of order 3 and then recover the nested pairwise links to build the motif ("fill in" the hyperedges, see Section 2.4.4); the sub-hypergraph is trivially connected since its nodes are part of the same hyperedge. Then, we can ignore all the higher-order interactions and focus only on the pairwise links, since we are interested in counting the frequency of the first two motifs of Figure 2.2. In this case, we can rely on ESU. This time, however, it will need to handle a lot fewer edges. Every time ESU returns an output, the triplet of nodes could have been counted already in the previous step (i.e., overlap between a pairwise motif and a hyperlink of order 3): in this case, the triplet is discarded. The first step has a complexity linear in the number of hyperedges of size 3, while the second step inherits the complexity of the ESU algorithm, resulting in a total complexity of $O(|E_3|) + O(N^k)$, where $|E_3|$ is the number of hyperedges of size 3, and k is the size of the subgraphs.

A formal description of the algorithm for higher-order motifs of order 3 is reported in Algorithm 2.2.

4-node motifs The algorithm for motifs of order 4 is similar, albeit there are some more details to take into account. One can still iterate on all the hyperedges of order 4, count the motifs by considering also the rich nested structures (one can observe that this time also hyperedges of order 3 can be nested), and discard all the 4-hyperedges. However, as a second step, one needs also to iterate over all the 3-hyperedges and consider all the possible neighbours; in fact, 3-hyperedges define only a sub-hypergraph with 3 nodes, while we are requesting 4 nodes. Neighbours can be listed by considering all the edges that add only 1 new node since 3 nodes are already fixed. The last step is to consider only pairwise interactions, and we rely again on ESU. Here, again, the first step has a complexity linear in the number of hyperedges of size 4. The second step has a complexity

Algorithm 2.2 Efficient algorithm: Counting higher-order motifs of order 3

Input: a hypergraph $\mathcal{H} = (V, E)$.
Output: distribution of the frequency of the motifs of order 3.

Let $\mathcal{M}$ be the motifs frequency hash map
for each hyperedge e of order 3 in E **do**
    $V^* \leftarrow$ vertices of e
    motif $\leftarrow$ sub-hypergraph induced by V^*
    Let $\mathcal{C}_m$ be the isomorphism class of *motif*
    $\mathcal{M}[\mathcal{C}_m] += 1$
end for
$G \leftarrow$ Discard all hyperedges of order 3 from $\mathcal{H}$
$\mathcal{S} \leftarrow \text{ESU}(G, 3)$
for each *subgraph* $= (V^*, E^*) \in \mathcal{S}$ **do**
    if V^* not already visited **then**
        Let $\mathcal{C}_m$ be the isomorphism class of *subgraph*
        $\mathcal{M}[\mathcal{C}_m] += 1$
    end if
end for

quadratic in the total number of hyperedges, being linear in the number of hyperedges of size 3 and then linear for each hyperedge to explore its neighborhood. Finally, the last step inherits the complexity of the ESU algorithm. Thus, the total computational complexity is $O(|E_4|) + O(|E_3| \cdot |E_2|) + O(N^k)$, where $|E_4|$ is the number of hyperedges of size 4, $|E_3|$ is the number of hyperedges of size 3, and k is the size of the subgraphs. A formal description of the algorithm for higher-order motifs of order 4 is reported in Algorithm 2.3.

2.4.4 Additional algorithm details

Counting higher-order motifs can be interpreted as the enumeration of all the possible connected sub-hypergraphs of size k, assigning each of them to an isomorphism class. An efficient way to assign an isomorphism class to a connected sub-hypergraph of size k (for small values of k) is relying on a hash map. One can generate and hash every possible pattern of higher-order interactions involving k nodes, with all the possible relabelings. Relabelings are important because the same sub-hypergraphs can be stored with different labels on the vertices. For example, we have 6 different patterns of higher-order interactions with 3 nodes, each with 3! possible relabeling; eventually, the hash map will contain $6 \cdot 3! = 36$ entries. One can use the hash map as a counter since each observed sub-hypergraph is a key. After having enumerated all the sub-hypergraphs, the final count of each motif is simply the sum of all the entries of the hash map that belong to the same isomorphism class. We show a summary of this process in Figure 2.4. Considering the

Algorithm 2.3 Efficient algorithm: Counting higher-order motifs of order 4

Input: a hypergraph $\mathcal{H} = (V, E)$.
Output: distribution of the frequency of the motifs of order 4.

Let $\mathcal{M}$ be the motifs frequency hash map
for each hyperedge e of order 4 in E **do**
    motif $\leftarrow$ sub-hypergraph induced by e
    Let $\mathcal{C}_m$ be the isomorphism class of *motif*
    $\mathcal{M}[\mathcal{C}_m]$ += 1
    Set vertices of *motif* as visited
end for
$\mathcal{H} \leftarrow$ Discard all hyperedges of order 4 from $\mathcal{H}$
for each hyperedge e of order 3 in E **do**
    Let $\mathcal{E}$ be the set of hyperedges adjacent to e
    for each hyperedge e_i in $\mathcal{E}$ **do**
        if $|e \cup e_i| = 4$ and $e \cup e_i$ not already visited **then**
            motif $\leftarrow$ sub-hypergraph induced by $e \cup e_i$
            Let $\mathcal{C}_m$ be the isomorphism class of *motif*
            $\mathcal{M}[\mathcal{C}_m]$ += 1
            Set vertices of *motif* as visited
        end if
    end for
end for
$\mathcal{H} \leftarrow$ Discard all hyperedges of order 3 from $\mathcal{H}$
$\mathcal{S} \leftarrow \text{ESU}(\mathcal{H}, 4)$
for each *subgraph* $= (V^*, E^*) \in \mathcal{S}$ **do**
    if V^* not already visited **then**
        Let $\mathcal{C}_m$ be the isomorphism class of *subgraph*
        $\mathcal{M}[\mathcal{C}_m]$ += 1
        Set V^* as visited
    end if
end for

sizes of the sub-hypergraphs involved, we can assume that this process incurs a constant time cost.

Another important routine in our algorithms is the construction of vertex-induced sub-hypergraphs. Given a set of vertices V', we are interested in querying the set of all the hyperedges to extract those who have all their endpoints in V'. This is what we referred to as "filling in" a set of vertices in the previous sections. For our specific case, this problem is efficiently solvable relying again on hash maps as follows. We can hash every hyperedge of a hypergraph: this ensures that we are able to check the existence of a hyperedge in constant time. Since we are only interested in solving this problem for a query set of vertices of size 3 or 4, we can easily generate all the possible 2^3 or 2^4 subsets of vertices (we can also ignore the empty set and the singletons) and check in constant time if each subset is an existing hyperedge. We show a summary of this process in Figure 2.4. All in all, given that we are interested in very small sets of vertices, we can construct vertex-induced sub-hypergraphs in constant time.

2.4.5 Sampling method for motifs in large scale hypergraphs

Scalability is a persistent issue for exact motif discovery algorithms. Motif analysis has a number of real-world applications that require handling vast datasets. However, exact algorithms for motif discovery quickly become intractable for realistic inputs and motif sizes. To address this complexity, we propose an approximated method based on hyperedge sampling.

Algorithm 2.4 samples with replacement S times a hyperedge e and enumerates all the connected sub-hypergraphs with a given number of nodes and containing e. The number of samples S controls the quality of the approximated results. However, directly sampling hyperedges from the hypergraphs leads to unreliable results. The distribution of hyperedge sizes is non-uniform, causing the algorithm to often sample hyperedges of size 2 while seldom sampling those of size 4. This skews the estimation of specific sub-hypergraph patterns. To mitigate this, we employ stratified sampling, segmenting the sampling process to guarantee a balanced consideration of hyperedges across different sizes. Let S_k be the number of samples assigned to hyperedges of size k, such that $S = \sum_k S_k$. We estimate appropriate values for S_k for every k empirically, exhaustively searching among different combinations of values and selecting those that maximize a defined quality function (see Appendix C.1).

The sampling algorithm proceeds in a way similar to the exact method (therefore we avoid explicitly repeating some details in the pseudocode). If we target the discovery of motifs

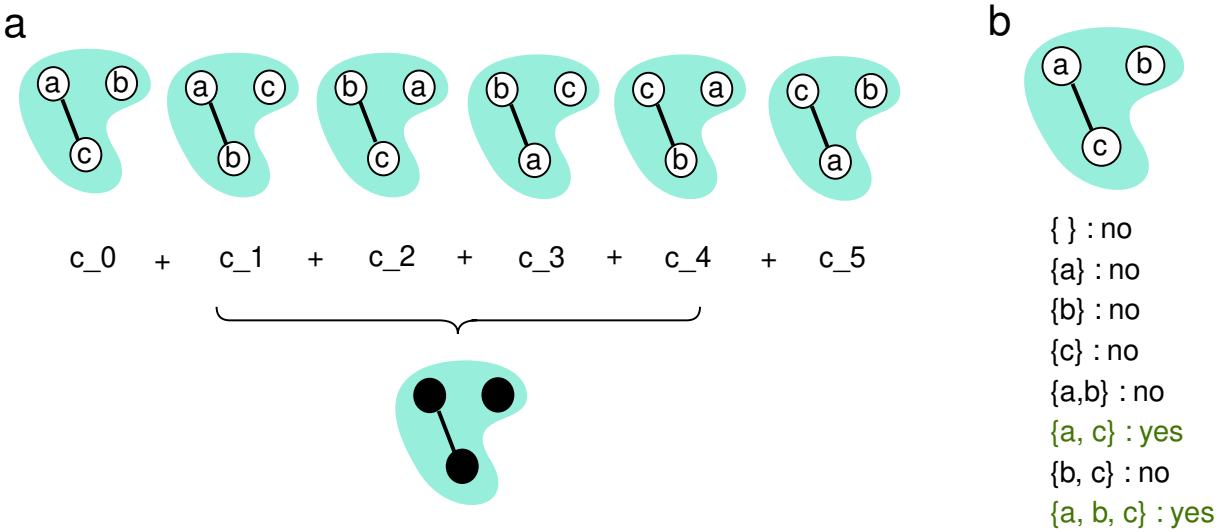

Figure 2.4: a) On the left, we show how to efficiently solve the problem of hypergraph isomorphism for small hypergraphs. We generate and hash every possible pattern of higher-order interactions involving k nodes with all the corresponding relabelings. Every observed sub-hypergraph will be equivalent to one and only one of the entries of the hash map. The final count of each motif is the sum of all the entries of the hash map that belong to the same isomorphism class. b) On the right, we show how to construct vertex-induced sub-hypergraphs efficiently. As a preprocessing step, we hash every hyperedge in a hypergraph, allowing us to check for their existence in constant time. For a query set of 3 or 4 vertices, we generate all the possible 2^3 or 2^4 subsets of the query set and check in constant time if each subset is an existing hyperedge. Every time a subset is found to exist, we add it to the sub-hypergraph induced by the query set.

of size k, then for all the sampled hyperedges of size k we have all the necessary vertices to build a target sub-hypergraph. No exploration of the neighbourhood is further required to add new nodes to the pattern. The complexity of this step is linear in the number of samples S_k. Then, for all the sampled hyperedges of size less than k, some exploration of possibly different levels of the neighbourhoods is required. The complexity of each of these steps is linear in the number of samples S_k multiplied, for each level of exploration, by a factor linear in the number of hyperedges. Moreover, for each pattern, the previously mentioned process of "filling in" the hyperedges is repeated to build vertex-induced sub-hypergraphs and count the right instances of the motifs. Again, these routines take constant time. Overall, the computational complexity remains the same as in the exact case, but the number of edges considered is reduced, depending on the chosen sampling parameter.

In order to estimate the exact count for each motif, the algorithm multiplies the observed count by a correction factor given by the probability of sampling a certain motif, as reported in the pseudocode. To simplify the computation of the correction factor, the algorithm discards all sub-hypergraphs encountered during the exploration of the neighbourhood of a hyperedge e that contains at least one hyperedge with a higher cardinality than e. In other words, a pattern of sub-hypergraph is only considered when the hyperedge of maximal cardinality is sampled. Given this approach, it is straightforward to prove that the estimator is unbiased.

2.5 The microscale organization of real-world hypergraphs

For our motif analysis of real-world higher-order systems, we collected a number of freely available networked datasets. The datasets [30, 86–102] come from a variety of domains: sociology (proximity contacts, votes), technology (e-mails), biology (gene/disease, drugs) and co-authorship. Each dataset has been manually tagged and associated with a specific domain. The description of each dataset is reported in Appendix A.1. In some datasets, higher-order structures are naturally encoded as hyperedges (e.g. three authors collaborating on the same paper), in others, we infer higher-order structures from pairwise interactions (e.g. for face-to-face interactions recorded over time, we promote cliques of size k to hyperedges of order k if the corresponding three dyadic encounters happened at the same time. We note that the choice of the specific time-window for aggregation does not affect our results, as presented in Appendix B.1.1.

Algorithm 2.4 Sampling algorithm: Counting higher-order motifs of order 4

Input: a hypergraph $\mathcal{H} = (V, E)$, number of hyperedge to sample S.
Output: approximated distribution of the frequency of the motifs of order 4.

Let $\mathcal{M}$ be the motifs frequency hash map
Let E_k represent the set of hyperedges of $\mathcal{H}$ of size k
Let S_k be the number of samples assigned to hyperedges of size k (ensure that the sum of the S_k for every possible k is equal to S)
sampled_ edges $\leftarrow$ sample S_2 hyperedges from E_2
sampled_ edges $\leftarrow$ *sampled_ edges* $\cup$ sample S_3 hyperedges from E_3
sampled_ edges $\leftarrow$ *sampled_ edges* $\cup$ sample S_4 hyperedges from E_4
for each hyperedge e in *sampled_ edges* **do**
    Let $\mathcal{S}$ be the set of connected sub-hypergraphs of $\mathcal{H}$ containing e
    for each *sub-hypergraph* in $\mathcal{S}$ **do**
        if *sub-hypergraph* contains a hyperedge of cardinality bigger than e **then**
            continue
        end if
        Let $\mathcal{C}_m$ be the isomorphism class of *sub-hypergraph*
        $\mathcal{M}[\mathcal{C}_m]$ += 1
    end for
end for
for each motif m with count c in $\mathcal{M}$ **do**
    Let *maxcard* be the maximum size of the hyperedges in m
    Let *countmax* be the number of the hyperedges in m of size *maxcard*
    $c \leftarrow \frac{c \cdot |E_{maxcard}|}{S_{maxcard} \cdot countmax}$
end for

2.5.1 Abundance and significance profile

The over- and under-expression measures of each higher-order motif (abundance with respect to a null model) in a hypergraph are concatenated in a Significance Profile that constitutes a fingerprint of the local structure of the network.

The abundance Δ_i of each motif i relative to random networks [61] is defined as:

$$\Delta_i = \frac{N\text{real}_i - \langle N\text{rand}_i \rangle}{N\text{real}_i + \langle N\text{rand}_i \rangle + \epsilon} \tag{2.1}$$

Following [61], we set $\epsilon = 4$. As a null model, we use the configuration model proposed by Chodrow [39]. We sample from the configuration model $n = 100$ times and compute the frequencies of the higher-order motifs in each sample.

We define the Significance Profile (SP) of a network as the vector of Δ_i normalized to length 1,

$$\text{SP} = \frac{\Delta_i}{\sqrt{\sum \Delta_i^2}} \tag{2.2}$$

2.5.2 Motifs of order 3

We start our motif analysis of empirical hypergraphs by characterizing their local connectivity at the smallest scale, with higher-order motifs of order 3. After having computed the SPs of all the datasets, the first question one could ask is how hypergraphs from different domains differ on average in their SPs. We compute the SPs of a domain by grouping and averaging the SPs of all networks that belong to it (more information about the disagreggated SPs can be found in Appendix B.1.3). The analysis of the higher-order profiles of order 3 of each domain highlights the relative structural importance of certain patterns of higher-order interactions (Fig. 2.5a). The pairwise triangle II appears to be a highly over-expressed motif in all the domains, whereas the greatest differences across domains emerge from motifs which involve a 3-hyperedge and at least one dyadic edge. In the social and technological domains, the motif VI made by a 3-hyperedge and a triangle of dyadic edges is highly over-expressed, suggesting that entities interacting in groups also tend to interact individually. In co-authorship networks, the most over-expressed motifs are IV and V, which involve a 3-hyperedge and one or two dyadic edges, indicating that in these domains there might be a hierarchical structure that prevents all nodes from interacting equally in pairs, as in the case of a research leader that co-authors papers

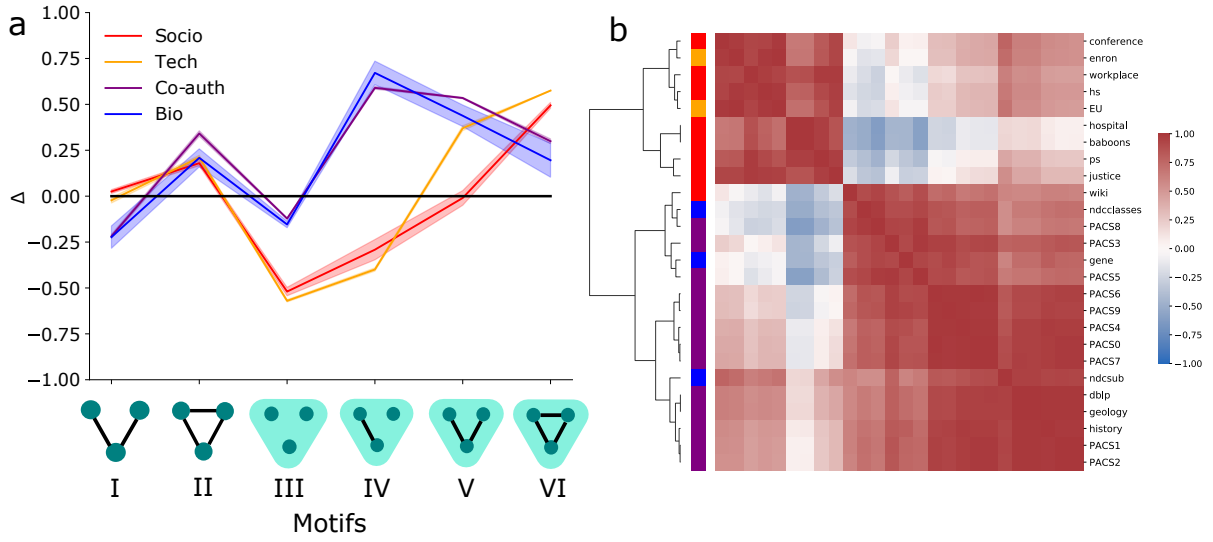

Figure 2.5: **A higher-order fingerprint for hypergraphs at the network microscale. a)** Significance Profiles (SP) of hypergraphs from higher-order motifs of order 3 (labelled I to VI). Δ is the abundance of each motif relative to random networks. Over-expressed higher-order motifs are associated with specific functionalities of the system. To simplify the plot, we averaged and grouped higher-order motif profiles of networks from the same domain. For each domain, we represent the mean of the respective higher-order motif profiles with a solid line and the standard error of the mean with a shaded area. **b)** Correlation matrix of the investigated datasets computed on the SPs. SPs of networks from similar domains display a positive correlation. We identify two large higher-order families of hypergraphs, characterized by distinct higher-order connectivity patterns at the local scale. Each row of the correlation matrix is labeled with different colors depending on the domain of the respective dataset: red for the social domain, orange for e-mails, purple for the co-authorship domain and blue for the biological domain. Moreover, we show the clustering tree computed by applying a hierarchical clustering algorithm on the significance profiles, considering correlation as a measure of similarity. The clustering tree highlights the hierarchical organization of the emerged clusters. In the correlation matrix, red squares represent high positive correlation while blue squares represent high negative correlation.

with students and postdocs while the latter do not co-authors papers without the former. A similar motif is also found to be over-expressed in biological systems. Moreover, SPs allow also to analyze anti-motifs, i.e. motifs that are highly under-expressed. An anti-motif in the social and technological domains is III, the 3-hyperedge without any dyadic interaction, indicating that it is unlikely that an interaction in group is not followed or preceded by any pairwise interaction. The biological and co-authorship domain do not display any anti-motif.

Another interesting question is whether the domain categorization naturally emerges from individually clustering the SPs of all the empirical hypergraphs. We perform a hierarchical cluster analysis considering the pairwise correlation between the distributions of the occurrences of the higher-order motifs for each dataset as (the inverse of) a distance (Fig. 2.5b). The analysis shows the emergence of two main clusters, i.e. families of higher-order networks that share similar patterns of higher-order interactions at the microscale. The clusters, here inferred in a purely data-driven manner, reproduce the partitions of domains displayed in Fig. 2.5a (social and technological datasets in a cluster, biological and co-authorship ones in the other), offering a more nuanced view on the similarity across datasets.

2.5.3 Motifs of order 4

In the previous section we have systematically investigated the smallest higher-order motifs. The number of possible patterns of higher-order interactions involving 4 nodes is significantly higher than the corresponding with 3 nodes, as it grows from 6 to 171. Despite the difficulties associated to this increase, analysing higher-order motifs of order 4 provides more nuanced information about the local structure of networks compared to 3-motifs.

In Fig. 2.6a, we group together similar domains based on the analysis in the previous section showing the average of their SPs with the higher-order motifs of order 4. The order of motifs along the x-axis maximizes the visual difference in SPs across clusters. On the left-end of the x-axis, we find motifs that are highly over-expressed in the Bio/Co-auth domain, while they are under-expressed in the Socio/Tech domain. Conversely, on the right-end of the x-axis we find motifs that are over-expressed in the Socio/Tech domain, while not characteristic for the other domain. This observation suggests that both the extremes of the x-axis carry information about the structural differences among the clusters.

The richer structural information captured by the higher-order motifs of order 4 compared

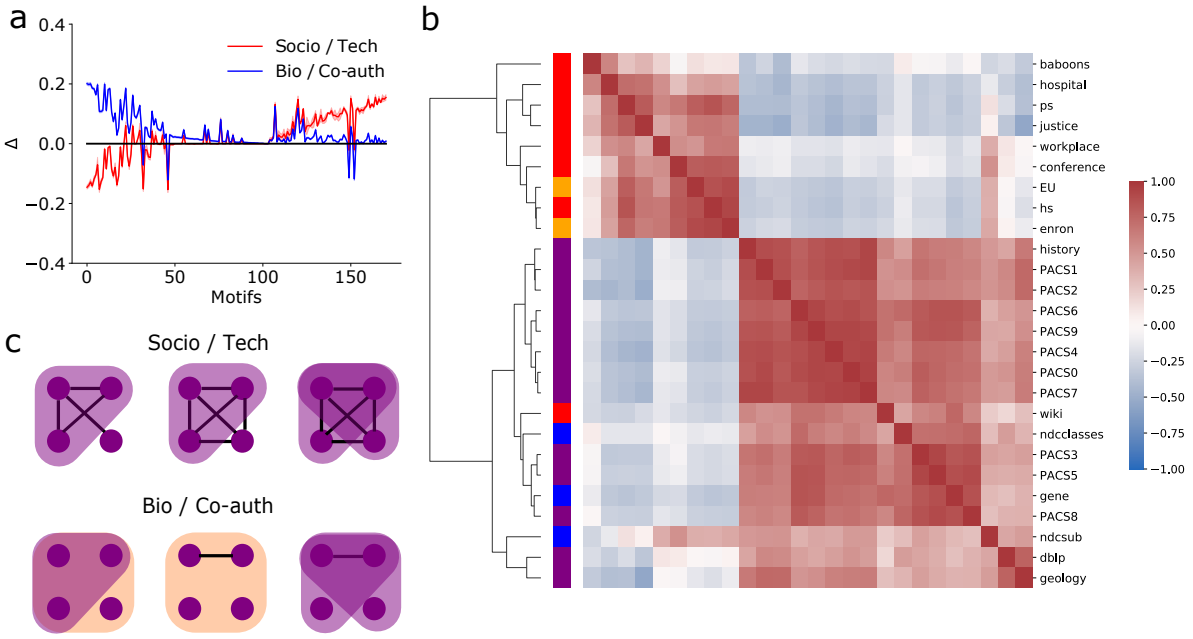

Figure 2.6: **Analyzing the local structure of hypergraphs via higher-order motifs of order 4. a)** Significance Profiles (SP) of hypergraphs from higher-order motifs of order 4. Δ is the abundance of each motif relative to random networks. SPs are much more complex due to the increase in the number of considered patterns of higher-order interactions. We group and average the SPs of networks from the same higher-order family (i.e. Socio/Tech and bio/Co-auth) and sort the motifs on the x-axis based on their ability to discriminate the two higher-order families. Distinct characteristic higher-order motifs of order 4 are associated to the two classes of networks. The shaded area represents the standard error of the mean. If the shaded area is not visible, it is of the same size of the line thickness. **b)** Correlation matrix of the investigated datasets computed on SPs of order 4. The matrix provides richer information than its equivalent at order 3 on the local structure of networks: the two big clusters emerge again but are better separated, and display a richer intra-cluster hierarchical structure. Each row of the correlation matrix is labeled with different colors depending on the domain of the respective dataset: red for the social domain, orange for e-mails, purple for the co-authorship domain and blue for the biological domain. Moreover, we show the clustering tree computed by applying a hierarchical clustering algorithm on the significance profiles, considering correlation as a measure of similarity. With respect to the analysis with higher-order motifs of order 3, the clustering tree highlights a better separation between the two big clusters, as well as a richer intra-cluster hierarchical organization. In the correlation matrix, red squares represent a high correlation while blue squares represent a low correlation. **c)** The six most representative higher-order motifs from the two clusters. Purple shaded triangles and orange shaded squares represent respectively higher-order interactions of size 3 and 4, whereas black lines represent pairwise interactions.

to their counterparts of order 3 is highlighted in the clustering analysis (Fig. 2.6b). When focusing on the two main clusters, the results are comparable with the previous cluster analysis. However, a richer hierarchical intra-cluster organization naturally emerges, as well as a better separation between the two clusters (See Appendix B.1.2).

Finally, we characterize the Socio/Tech and the Bio/Co-auth clusters by means of their most over-expressed, and therefore most representative, higher-order motifs of order 4 (Fig. 2.6c). The Socio/Tech domain shows an over-expression of structures involving more lower-order nested relations (e.g. dyadic links), while the Bio/Co-auth domain displays a preference towards less relations but of higher-order. This pattern might be caused by the fact that people interacting in groups are likely to interact also in single pairs, therefore it is plausible that group interactions in the Socio/Tech domain are supported by a large number of lower-level interactions. On the other hand, people tend to write papers in large groups and tend to maintain the same research group over time, with few additions or removals. Therefore patterns involving only dyadic relations are penalized. For a more in-depth description of the most over- and under-expressed higher-order motifs of order 4, we refer to Appendix B.1.4.

2.5.4 Motifs of order 5

Exact counting algorithms are suited only for the extraction of motifs of order 3 and 4. Instead, the previously proposed approximated method not only speeds up motif analysis for large datasets but also allows for the study of larger patterns of interactions. Here we use such a method to characterize two real-world hypergraphs, namely `history` and `high school` in terms of their higher-order motifs of order 5. We use the same statistical evaluation methods proposed before, i.e., we consider the relative abundance of each motif with respect to a configuration model [39]. We run the approximated algorithm also on the randomized instances. In Figure 2.7 we show the most over-expressed higher-order motifs of order 5 in both the real-world hypergraphs. We can notice how also at this scale we still observe characteristic patterns of co-authorship data (low number of interactions of large average size) and face-to-face data (high number of interactions of small average size). This is in line with the previous results. In the following sections, we will provide a more thoughtful evaluation of the accuracy of this sampling algorithm.

2.5.5 Nested organization of higher-order interactions

As motif analysis becomes highly computationally demanding when the order increases, in the following we focus on characterizing the nested structures of large hyperedges.

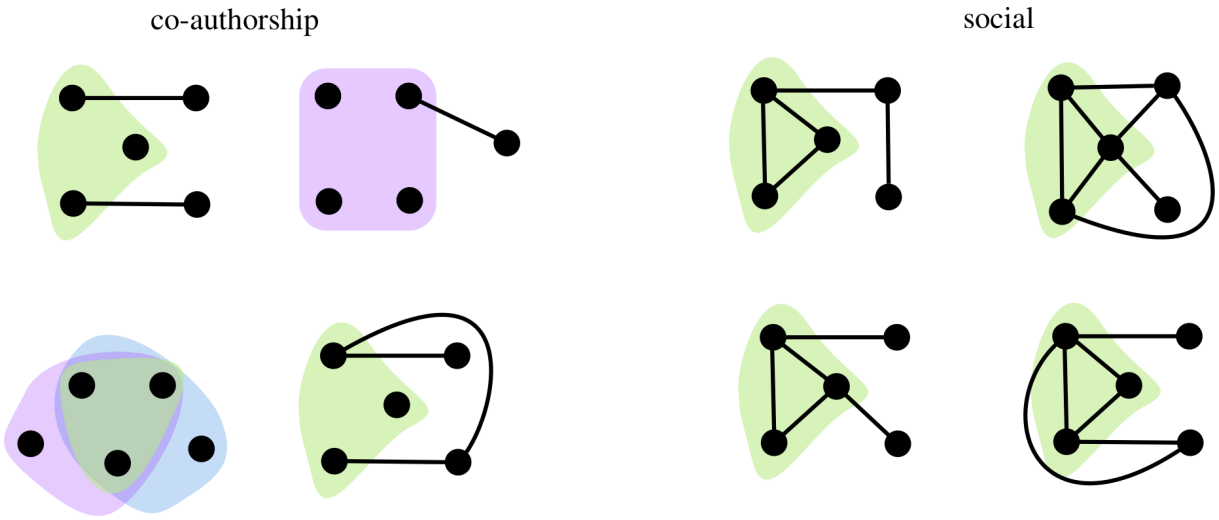

Figure 2.7: Over-expressed patterns of higher-order interactions highlight structural principles of the different domains.

This means that, instead of counting the exact frequency of each pattern of higher-order interactions, we settle for extracting statistics on the hierarchical structures of interactions inside hyperedges of any size. We define the nested structure of a large hyperedge h as the collection of hyperedges existing on a subset of the nodes of h, and extract statistics on the nested structure of hyperedges of any size. The advantage of this approach is that it still provides information about the local structure of sub-modules of a network, while its computational complexity is only linear in the number of hyperedges in the hypergraph.

First, we consider the average number of edges in the nested structures of hyperedges of different sizes (Fig. 2.8a). The networks are grouped according to their domain. While biological and co-authorship networks do not display evident differences in the number of nested edges with the growth of the hyperedge size, social and technological networks show a clear growing trend with a change of slope after orders 5 and 6.

In order to complement this information, we looked at how the mean size of the nested edges changes with the growth of the size of the analyzed hyperedges (Fig. 2.8b). In this case, all the domains show a growing trend, with biological and co-authorship networks displaying a faster growth. Thus, while social and technological networks tend to have more edges in the nested structure of their large hyperedges, they tend to be of small size. Biological and co-authorship networks, instead, shows an opposite behavior. All in all, this suggests that, in agreement with our previous findings, also at higher scales Socio/Tech network motifs are systematically more nested.

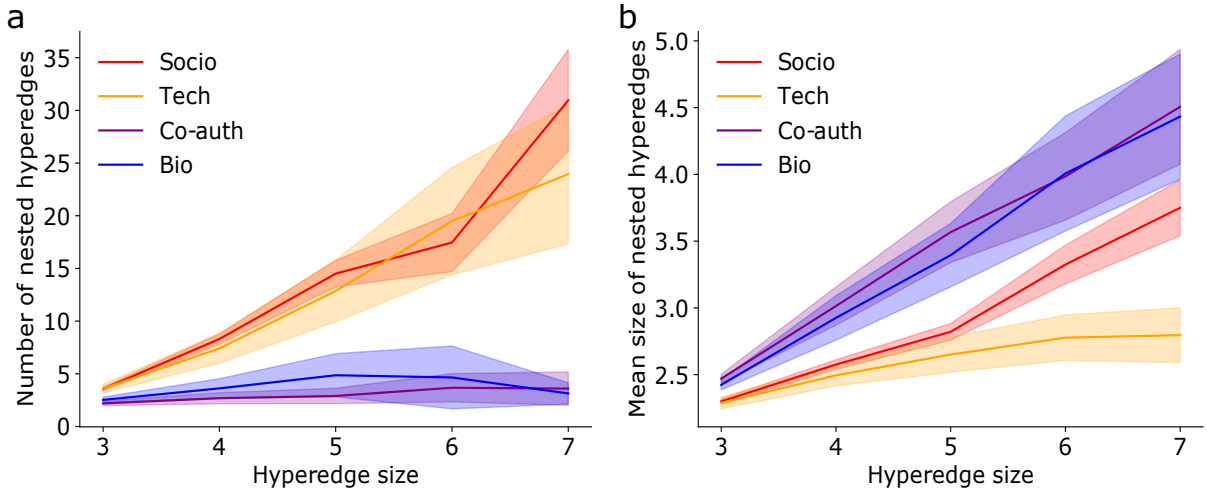

Figure 2.8: **Nested organization of group interactions.** Different higher-order families of hypergraphs can display very different hierarchical organization of their higher-order interactions. **a)** Mean number of hyperedges in the nested structure of large hyperedges as a function of their size. Biological and co-authorship networks display a static behavior, while social and technological networks show a clear increasing richness of the nested hierarchical structures of the hyperedges. **b)** Mean average size of the hyperedges in the nested structure of large hyperedges as a function of their size. All the domains show a linear growing trend, however biological and co-authorship networks grow faster. All in all, Socio/Tech networks tend to have a lot of small-size edges in the nested structure of their hyperedges. The Bio/Co-auth domain, instead, tend to prefer few large-size edges. In both panels, the shaded area represents the standard deviation.

2.5.6 Higher-order motifs and reinforcement

In order to understand if and how the occurrence of nested dyadic interactions affects the strength of group interactions, we investigate how much the weight of each hyperedge (i.e. the number of times each group interaction occurs) is correlated with the number of nested pairwise links. We find that a positive trend emerges, indicating the existence of a correlation between a rich nested pairwise structure and the weight of a hyperedge (Fig. 2.9a). We dubbed this phenomenon, similar to the one highlighted in Ref. [22] for multilayer networks, as higher-order structural reinforcement.

Moreover, we used the metadata about personal relationships between students recorded in the High School dataset from SocioPatterns to understand if a similar reinforcing behavior is observed in the presence of friendship interactions between individuals. Friendship data has been collected in two ways, from Facebook accounts and through a questionnaire. In the first case, two students are always reciprocally friends, while in the second case a friendship can be unreciprocated. In Fig. 2.9b we analyze the relationship between the average number of friends (both on Facebook and by questionnaire) and the topology of the different motifs in the proximity hypergraph. Our results show that the higher the number of pairwise interactions between students that interact in hyperedges of size three, the higher will be the number of friends in the group, further suggesting the existence of reinforcement mechanisms.

2.6 Performance evaluation of mining algorithms

In this section, we turn our attention to the computational aspects of the problem and assess the improvement in the performance of the algorithms for higher-order motif discovery when (i) exploiting higher-order structures instead of applying classic methods on the hypergraph projection and (ii) approximating motif frequency. Besides the evaluation of the performance, we also study the accuracy of the sampling algorithm and exploit sampling methods to study higher-order motifs of order 5. All the experiments have been carried out on a machine with an 8-core (2.2GHz) Intel Xeon CPU and 94GB of RAM, running Ubuntu 20.04.4 LTS. The algorithms presented in this paper are implemented in Python3. The code is publicly available [103]. Moreover, all the algorithms presented in this work are included in the Python library Hypergraphx for higher-order network analysis [3].

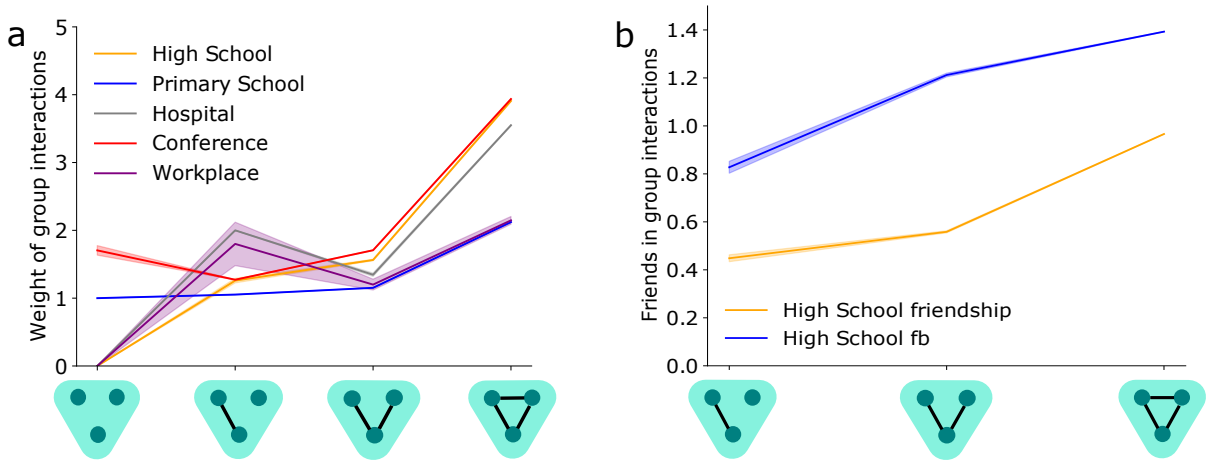

Figure 2.9: **Structural reinforcement.** A rich supporting nested structure of pairwise links makes group interactions stronger. In both panels, the stronger levels of connectivity are observed when the number of dyadic interactions increases. **a)** Mean weight of each group interaction (i.e. the number of times each group interaction occurs) as a function of the number of its nested pairwise links. **b)** Mean number of friends (certified by a Facebook friendship or by a questionnaire) in group interactions as a function of the number of their nested pairwise links. In both panels, the shaded area represents the standard error of the mean.

2.6.1 Datasets

For the performance evaluation, we collected a representative subset of the real-world datasets used in the previous sections. In particular, we used datasets from different domains, describing face-to-face interactions, co-authorship relations and e-mail communications. The summary statistics of the datasets are reported in Table A.1. The datasets, as well as the preprocessing scripts, are publicly available [103].

2.6.2 Running time

In Table 2.2 we compare our exact algorithm for higher-order motif discovery against the baseline algorithm in terms of their execution running time. We show that exploiting directly higher-order structures speeds up the computations. The efficient algorithm outperforms the baseline in every dataset. Moreover, it is worth mentioning that the analysis with motifs of order 4 of `dblp` with the baseline algorithm was not feasible in a reasonable amount of time. The larger gains are observed in co-authorship data. Co-authorship systems are proven to display a nested structure of hyperedges made up of a small number of hyperedges of large average size [104]. In fact, these kinds of systems are the ideal

Table 2.1: Summary statistics of the datasets considered for our experiments. Each higher-order network is described by the domain, the number of nodes, and the total number of hyperedges of size 2, 3, 4 and 5.

Dataset	Domain	N	E_2	E_3	E_4	E_5
`hs`	`proximity`	327	5498	2091	222	7
`ps`	`proximity`	242	7748	4600	347	9
`EU`	`e-mail`	998	12753	4938	2294	1359
`dblp`	`co-auth`	1924991	693363	667291	419434	205970
`history`	`co-auth`	1014734	160885	47423	19120	8775
`geology`	`co-auth`	1256385	275736	227950	159509	99140

Table 2.2: Comparison of the running time (s) of the exact algorithms with motifs of order 3 and 4.

Dataset	Base-3	Eff-3	Base-4	Eff-4
`hs`	7	**5**	362	**230**
`ps`	25	**18**	1920	**1339**
`EU`	44	**29**	5286	**2757**
`dblp`	1185	**134**	$>$ 24h	**2885**
`history`	42	**19**	4591	**526**
`geology`	207	**36**	32810	**475**

scenario for our algorithm. We can notice that the gains are not as noticeable in social datasets, which tend to be governed by dense patterns of lower-order interactions [104].

In Table 2.3 we show the execution running time in seconds of the sampling algorithm on the different datasets with multiple values of S, i.e., the parameter that controls the number of samples. The different size scales of the co-authorship and social datasets require different sample sizes to achieve results of comparable quality. Since the analysis of the motifs of order 3 was already easily doable, we consider only the task of motif discovery of order 4. We show that hyperedge sampling dramatically improves performance. As expected, the parameter S heavily affects the running time. As always, there is a trade-off between the accuracy of the results (higher values of S lead to more accurate estimates) and the execution running time.

2.6.3 Accuracy of sampling method

Besides evaluating the running time of the sampling method, it is also important to assess the output quality of the estimates compared to exact higher-order motif profiles. We compute motif profiles [104] comparing the observed frequencies of the motifs with those on a null model [39] to assess their statistical significance (we sample $N = 10$ times from the configuration model).

We evaluate the quality of the estimated motif profiles in terms of:

- the *Pearson's correlation coefficient* ρ between the estimated and the exact higher-order motif profiles. The coefficient assigns values close to 1 to profiles in strong agreement and values close to -1 to profiles in strong disagreement.
- the *Maximum Absolute Error (MaxAE)* in estimating higher-order motif profiles.
- the *Mean Absolute Error (MAE)* in estimating higher-order motif profiles.

In Table 2.3, we show that we obtain good results even with a small number of samples with respect to the total size of the hypergraphs. The measures of output quality improve with increases in S. A good trade-off between the output quality, S and the execution running time will be critical for real-world applications. The measures of output quality are averaged across 10 repetitions for every value of S.

A second evaluation metric is the correlation matrix of the motif significance profiles [60, 104] of the different real-world hypergraphs. In Figure 2.10, the correlation matrix shows the emergence of two "superfamilies" of real-world hypergraphs, in a way similar to the previous section. Clustering tends to separate social and co-authorship data. This further

Table 2.3: Hyperedge sampling dramatically improves the performance with respect to the exact algorithm. The execution running time of the approximated algorithm heavily depends on the choice of the sample size S. The correlation coefficient ρ between the estimated and the exact motif profiles, the maximum absolute error MaxAE and the mean average error MAE improve with increases in the number of samples S. Due to their different size scale, co-authorship and social datasets require different sample sizes to achieve comparable results. We obtain reasonable results even with a very limited number of samples.

Dataset	Exact exec. time (s)	S	Approx. exec. time (s)	ρ	MaxAE	MAE
`hs`	230	100	3	.914	.151	.015
		250	8	.953	.122	.011
		500	16	.978	.096	.007
		1K	29	.987	.065	.006
`ps`	1339	100	11	.918	.151	.017
		250	30	.950	.135	.012
		500	63	.977	.093	.008
		1K	118	.986	.071	.006
`EU`	2757	100	15	.804	.203	.028
		250	34	.887	.159	.020
		500	73	.923	.134	.016
		1K	144	.963	.098	.010
`dblp`	2885	1K	32	.495	.088	.055
		2.5K	41	.555	.088	.047
		5K	56	.610	.089	.038
		10K	85	.696	.087	.027
`history`	526	1K	5	.679	.176	.033
		2.5K	7	.804	.169	.022
		5K	11	.867	.170	.016
		10K	22	.913	.124	.012
`geology`	475	1K	12	.590	.107	.047
		2.5K	15	.661	.107	.039
		5K	19	.719	.106	.032
		10K	30	.784	.103	.024

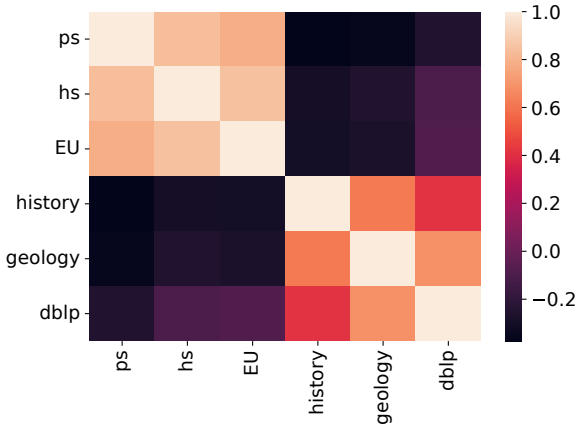

Figure 2.10: The correlation matrix of the significance profiles built with sampling methods ($S = 1000$ for co-authorship data and $S = 100$ for social data) highlights the emergence of two clusters that separate social and co-authorship data.

proves that sampling methods are still able to capture and highlight patterns of higher-order interactions that are probably linked to the functionalities of the networks.

2.7 Discussion

The framework of network motifs is widely recognized as a fundamental tool for the analysis of complex networks. Able to highlight local structural characteristics of networks and influence their dynamics, motifs can be considered the fundamental building blocks of networks, and have produced applications in a number of fields such as biology and social network analysis.

Modeling complex systems by means of hypergraphs have recently emerged as a fundamental tool in Network Science, prompting the question of how to identify and assess network motifs in the presence of higher-order interactions. With the aim of extracting the local fingerprint of hypergraphs, in this Chapter, we introduced the notion of higher-order network motifs, which are small, possibly overlapping patterns of higher-order interactions that are statistically over-expressed with respect to a null model. We proposed a combinatorial characterization of higher-order network motifs, as well as efficient algorithms to evaluate their statistical significance on empirical data. These tools allowed us to extract fingerprints of a variety of real-world systems by focusing on their characteristic patterns of higher-order interactions among small groups of nodes, showing the emergence of families of hypergraphs characterized by similar local structures. Moreover, we proposed a set of measures to study the nested structure of hyperedges

and provided evidence of a structural reinforcement mechanism that associates stronger weights of higher-order interactions to groups of nodes that interact more at the pairwise level.

Similarly to the case of traditional pairwise network motifs, we believe that higher-order network motifs can pave the way to applications in a number of domains, pushed by the growing awareness of the relevance of the higher-order nature of interactions in many real-world systems. All in all, our work highlights the informative power of higher-order motifs, providing an initial approach to extract higher-order fingerprints in hypergraphs at the network microscale.

3 | Hyperlink communities in higher-order networks

Many networks can be characterised by the presence of communities, where nodes within the same community share many more links than nodes in different communities [105]. In real-world systems, community structure is often hierarchical [106–109], with nested sub-units, and overlapping [110, 111], where nodes belong to multiple communities simultaneously. Link communities [112, 113] reconcile these properties by focusing on groups of related interactions rather than nodes. Indeed, while nodes can simultaneously belong to multiple groups (for instance, an individual might have a family, a circle of friends, and a team of colleagues), interactions typically occur for a singular purpose, such as two individuals connected by a shared interest or a family relation [113]. Recent work has begun to explore the mesoscale organization of hypergraphs [42–44, 114–118], however, this area remains relatively unexplored.

In this Chapter, we extend the concept of link communities to hypergraphs. This extension allows us to extract informative clusters of highly related hyperedges. We analyze the dendrograms obtained by applying hierarchical clustering to distance matrices among hyperedges across a variety of real-world data, showing that hyperlink communities naturally highlight the hierarchical and multiscale structure of higher-order networks. This approach also enables the extraction of overlapping node memberships and introduces higher-order network cartography for classifying nodes based on their interaction patterns and community participation, providing insights into individual roles across different social systems.

3.1 Hyperlink communities

The notion of hyperlink communities extends to hypergraphs the idea of describing the mesoscale structure of a system by grouping (higher-order) interactions instead of nodes. While atypical, this approach can describe the hierarchical organization of hyperedges (Fig. 3.1a) and node community overlap (Fig. 3.1b) as two aspects of the same phe-

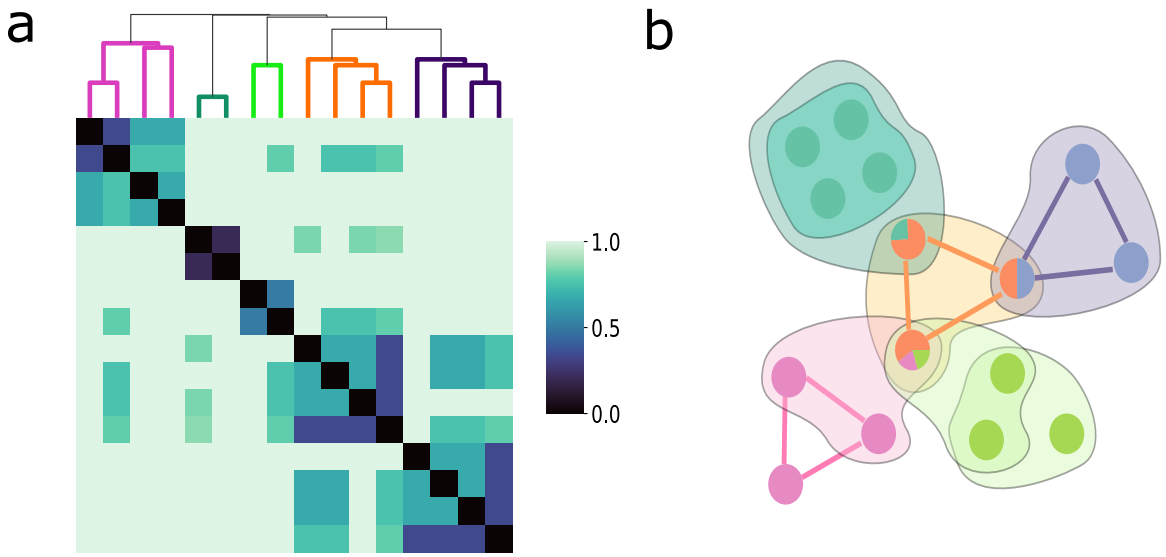

Figure 3.1: **Hyperlink communities and their properties.** Hyperlink communities group interactions to describe the mesoscale structure of a hypergraph. This approach is able to explain both the hierarchical organization of hyperedges and the overlap of communities among nodes. **a)** We perform hierarchical clustering on the hyperlinks of an observed hypergraph, considering their Jaccard distance. The output of such clustering is a dendrogram in which the leaves are the hyperlinks and the branches are the hyperlink communities. The dendrogram can be cut at different thresholds, each threshold potentially giving a meaningful community structure as output. **b)** After the cut, each hyperlink is uniquely assigned to a specific community. Nodes are then assigned to the set of communities to which the hyperlinks in which they are active belong. As a result, a single node may belong to multiple communities simultaneously.

nomenon. In a way similar to the seminal work on link communities [113], we define a distance measure between hyperlinks and perform hierarchical clustering on top of them to obtain hyperlink communities (Fig. 3.1a). Being each hyperlink encoded as a set of interacting nodes, a natural way to compute the distance between two hyperlinks A and B is to consider their Jaccard distance, defined as $J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$. To optimize the computation of distances, we precompute and cache sets of hyperedges with at least one common node. This strategy allows our distance algorithm to execute the more time-consuming steps of set-union and set-intersection only when strictly necessary. It is also worth noting that this step can be easily parallelized or distributed in the case of very large hypergraphs. We perform single-linkage hierarchical clustering on top of the distance matrix of the hyperlinks. The clustering procedure follows a bottom-up approach: it starts by assigning each hyperlink to its own cluster and then merges the clusters of hyperlinks with the smallest average distance until all the hyperlinks are part of a single cluster. The output of hierarchical clustering applied on such hyperlinks is a dendrogram. The dendrogram is a graphical representation of the hierarchical structure of the communities formed during the clustering process. In such a dendrogram, leaves are hyperlinks from the observed hypergraph and branches are hyperlink communities. Moreover, the height in the dendrogram of each branch provides additional information about the strength of the merged communities. For a formal description of the process of building the dendrogram, we refer to algorithm 3.1. The dendrogram can be cut at different heights, or thresholds, giving as output different meaningful community structures, typically revealing distinct multiscale organization of hyperlink communities.

Algorithm 3.1 Hierarchical clustering on hyperlinks

Input: a hypergraph $\mathcal{H} = (V, E)$
Output: the dendrogram of hyperlink communities
1: $\mathcal{D} \leftarrow$ matrix of pairwise distances of hyperlinks
2: clusters $\leftarrow$ each hyperlink is assigned to a singleton cluster ▷ The leaves of the dendrogram.
3: **while** number of clusters > 1 **do**
4: merge the clusters with the smallest distance ▷ The branches of the dendrogram. The height of a branch is the distance between the two merged clusters.
5: **end while**

While after a cut each hyperlink is uniquely assigned to a community, nodes inherit the community memberships of all the interactions they participate in. Nodes might participate simultaneously in several communities, although with different strengths. Therefore, another property of hyperlink communities is the ability to naturally extract overlapping communities of the nodes of a hypergraph (Fig. 3.1b).

Dataset	N	E	E_2	E_3	E_4	E_5	Domain
Phys-soc	26800	15311	3700	4135	2854	1613	Co-auth
NDC_classes	1161	1088	297	121	125	94	Bio
PACS3	33479	16977	2099	2105	2590	1169	Co-auth
ENRON	143	1512	809	317	138	63	E-mail
Primary school	242	12704	7748	4600	347	9	Proximity
High school	327	7818	5498	2091	222	7	Proximity
Hospital	75	1825	1108	657	58	2	Proximity
Baboons	13	231	78	142	11	0	Proximity

Table 3.1: Details of the real-world networked datasets considered for our experiments. Real-world hypergraphs from different domains are described by their number of nodes, total number of hyperedges and number of hyperedges of size 2, 3, 4 and 5.

In order to study real-world hypergraphs, we gathered a collection of freely available datasets of systems with group interactions. The datasets come from a variety of domains, including face-to-face proximity contacts in a primary school, in a high school, in a hospital and in a group of baboons [14, 30, 92, 102, 119], e-mail exchange (Enron) [30], biology (NDC classes, i.e., class labels applied to drugs) [30] and co-authorship in the physics area (PACS3: Atomic and Molecular Physics, Arxiv physics and society) [87]. More information about the datasets used in our experiments is reported in table 3.1.

The code of the experiments is freely available [120] and it is implemented as part of hypergraphx [3], our open source python library for higher-order network analysis.

3.2 Multiscale properties of hyperlink communities

As previously mentioned, the output of the hierarchical clustering algorithm applied to the distance matrix of the hyperlinks is a dendrogram in which the leaves are the hyperlinks, the branches are hyperlink communities, and the height of a branch encodes information about the strength of a hyperlink community. By cutting the dendrogram at different thresholds, we can obtain different community structures, each of which may provide valuable insights into the organization of the hypergraph. This feature is useful for performing analyses of a system at different scales. For example, in a time-varying system, a student might interact more with students sitting nearby during a lecture, with other students in the same class during a break, and with students from other classes during lunch. When aggregated over time, such a variety of patterns might lead to a complex multiscale organization of interactions. The shape of the hierarchical clustering dendrogram depends on the patterns of overlap between hyperlinks. To illustrate how

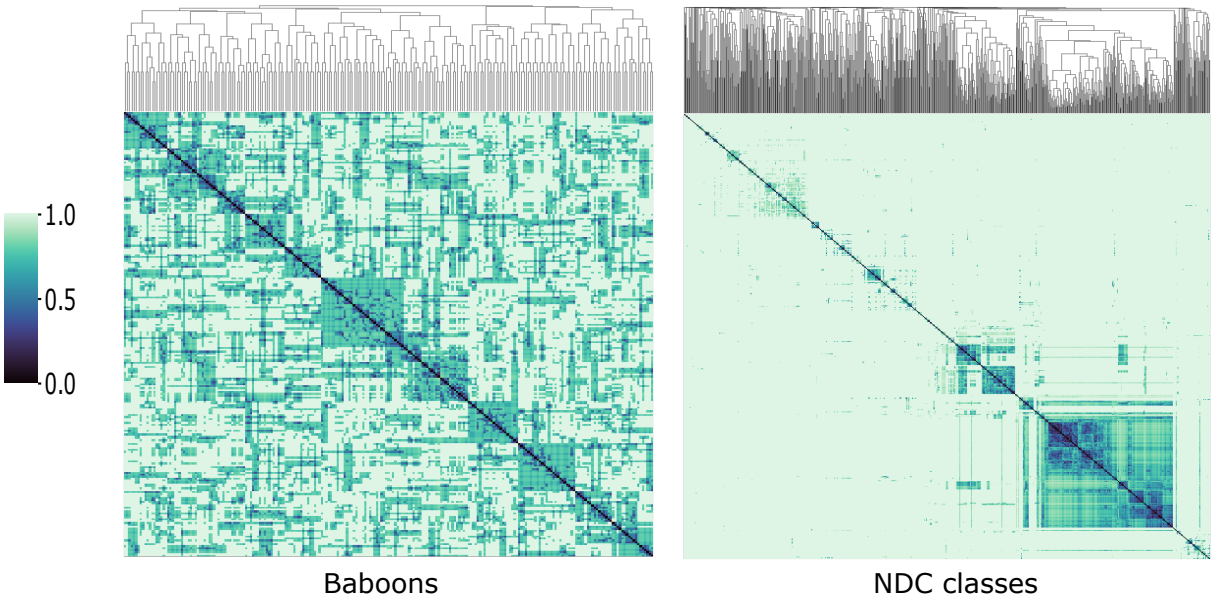

Figure 3.2: **Hierarchical clustering of hyperlinks in real-world hypergraphs.** We provide two examples of dendrograms (and their corresponding distance matrix) of hyperlink communities from real-world hypergraphs: one representing proximity group interactions among baboons, and the other representing affiliations between drugs and class labels applied to each drug. Hypergraphs can show very different hierarchies of hyperlinks, due to different statistics of their overlap distances. In particular, we identified two broader classes of real-world hypergraphs, of which these two examples are representative.

hypergraphs can have very different hierarchies of hyperlinks, depending on the distribution of their overlap distances, in Fig. 3 we show two examples of dendrograms (and their corresponding distance matrix) of hyperlink communities from real-world hypergraphs: one representing proximity group interactions among baboons and the other representing drugs and their associated class labels. In the following, we show that these two examples are indeed representative of two broader classes of real-world hypergraphs.

To describe and build a profile of the hierarchy of real-world hypergraphs, we can measure the number of hyperlink communities at different increasing thresholds. The specific scaling of the number of hyperlink communities can be interpreted as a fingerprint of the hierarchy of group interactions in real-world systems. In Fig. 3.4a, we show such profiles for several datasets from different domains, highlighting the emergence of different behaviours. Data on social proximity, such as contacts within hospitals and schools, reveals distinct spikes at certain thresholds (red lines). These spikes correlate with a prevalent recurrence of specific overlapping patterns among hyperlinks. This behaviour indicates that the dendrograms are hierarchically organized in a few, important distinct

topological levels. Instead, affiliation and co-authorship data show a smoother decreasing curve (blue lines). While such datasets approach 0 at differing rates, they do not exhibit significant spikes. This behaviour indicates that smoother hierarchical transitions are associated with the different system scales. Our multiscale analysis highlights the emergence of two families of real-world hypergraphs, distinguished by their scaling profiles. This characterization of hypergraphs is in agreement with purely local methods, such as motif analysis [1].

Statistics about the hyperlink communities obtained by cutting the dendrogram at different thresholds can vary significantly. Fig. 3.4b shows the distributions of the size of hyperlink communities at three different significant cuts for a subset of the selected datasets, with examples from the two identified classes of real-world hypergraphs. The community structure can change significantly across different scales. This demonstrates that not only do the dendrograms encode information about a hypergraph at multiple scales, but it is also valuable to analyze each level since they can provide distinct insights into the global organization of a system.

3.3 Overlapping communities at multiple scales

Previously we mentioned that by cutting at a certain threshold the dendrogram constructed by applying hierarchical clustering on the hyperlinks distance matrix, not only does the algorithm uniquely assign a community to each hyperlink, but also assigns multiple (possibly overlapping) communities to nodes. By varying the dendrogram-cutting threshold, we can extract overlapping communities across various scales. Having fixed a dendrogram-cutting threshold, for a node n, we can define the community membership vector v_n as:

$$v_n = \{c(e) : e \in E(n)\} \tag{3.1}$$

where $E(n)$ is the set of hyperlinks in which n participates, and the function $c(e)$ assigns a hyperlink to its community.

In the following, we study the community membership vectors of the nodes across datasets and scales, to quantify node community overlap in real-world hypergraphs.

In Fig. 3.5, we analyze the distributions of node community sizes (i.e., the number of nodes participating in a community) and node community memberships (i.e., the communities a node simultaneously participates in) at different cuts of the dendrogram. We find

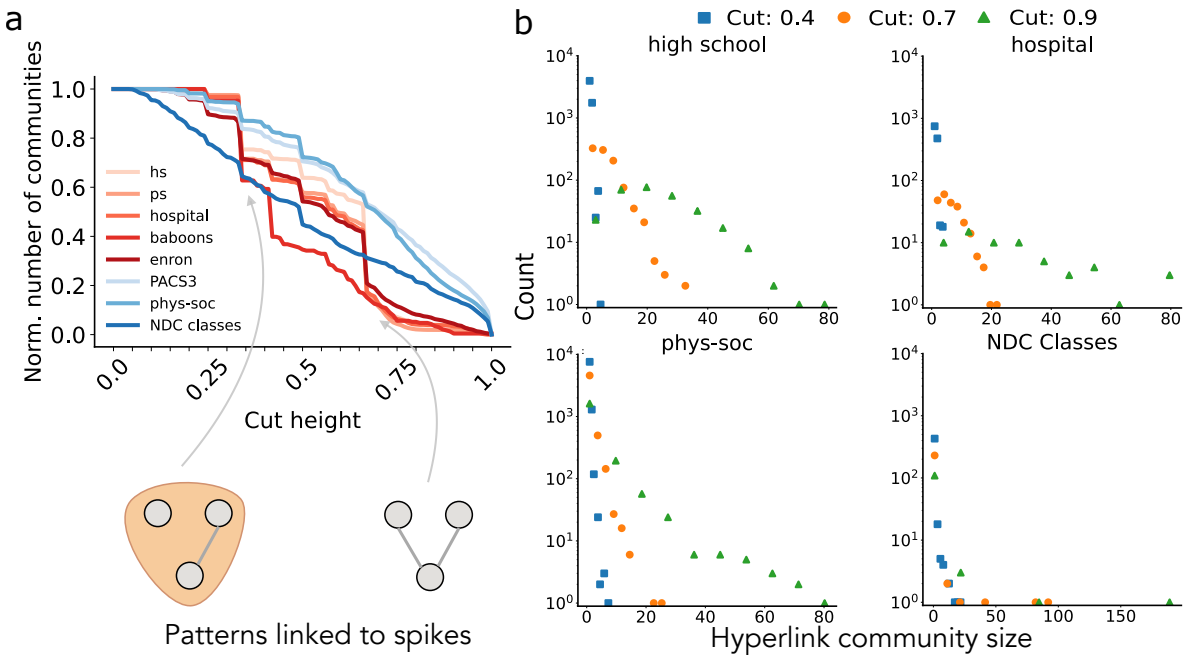

Figure 3.3: **Multiscale properties of higher-order networks.** Hierarchical clustering dendrograms can be cut at several thresholds, allowing for the extraction and analysis of hyperlink communities at multiple scales. **a)** The scaling of the number of hyperlink communities can be interpreted as a fingerprint of the hierarchical organization of group interactions in real-world systems. Due to the over-abundance of certain patterns of overlap between small group interactions social proximity data (red lines) show clear spikes in their curves. **b)** Evolution of the statistics of the hyperlink communities at different thresholds. Hyperlink community structures can change significantly across scales.

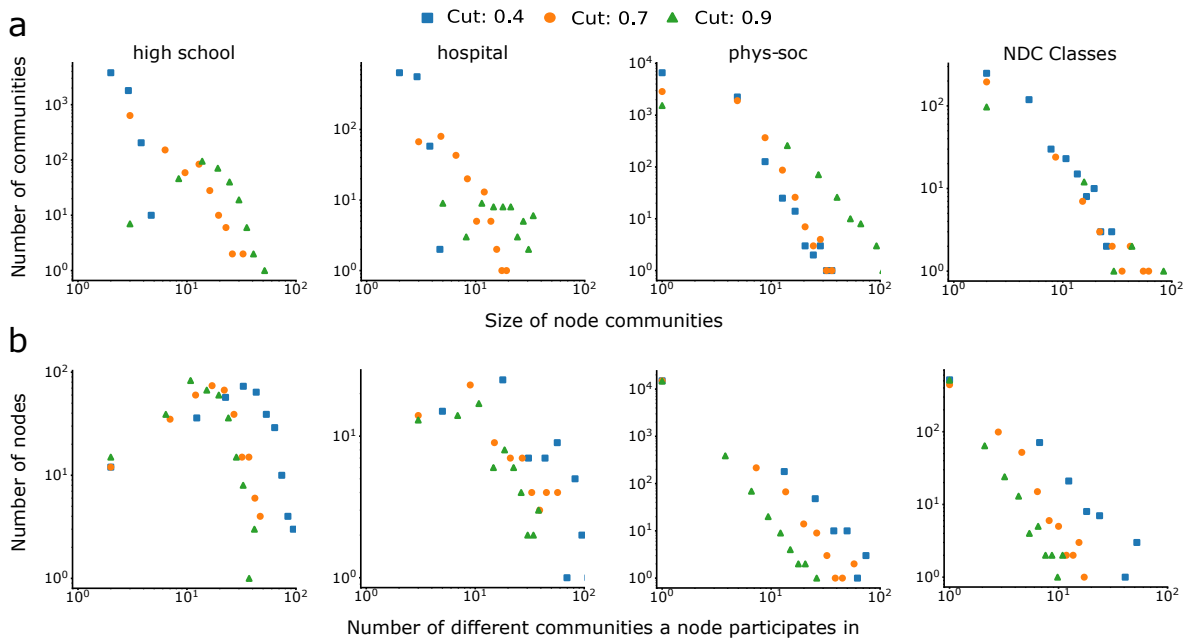

Figure 3.4: **Statistics of overlapping communities at multiple scales.** The distribution of node community sizes and node community memberships for several hypergraphs at three different dendrogram thresholds reveals the multiscale overlapping structure of real-world hypergraphs at their mesoscale. The hypergraphs show a wide range of community sizes, generally exponentially distributed, throughout the dendrogram. The distributions of community memberships per node show that nodes tend to participate simultaneously in more communities. This behaviour is consistent across scales. Proximity data has a more pervasive overlapping structure than the other datasets.

real-world hypergraphs to exhibit a wide spectrum of node community sizes, generally exponentially distributed. The distributions of community memberships per node at different cuts of the dendrogram remain in line across cut heights and show that nodes tend to consistently participate in more than one community at the same time. This suggests that real-world hypergraphs present overlapping structures at multiple scales. Moreover, the different shapes of the distributions highlight once again the different behaviour of the two families of datasets, with proximity data having a more pervasive overlapping structure.

In Fig. 3.6a, we provide a comparison between the overlapping communities extracted from real-world hypergraphs and the associated metadata. For the hospital dataset, describing proximity interactions among people inside a hospital, each node is characterized by metadata about their role (e.g., nurse). For the high-school and primary-school datasets, describing proximity interactions among students and teachers inside a primary school and a high school, we have information about the class of each individual. For each node n, we compute a binary community membership vector. In this vector, for a node n, entry i is equal to 1 if node n participates in the community i at the representative cut, and equal to 0 otherwise. To identify nodes with similar connectivity patterns, we compute pairwise similarities between binary vectors using the Jaccard similarity. For each dataset, we create the role-to-role similarity matrix $\mathcal{M}$ in which $\mathcal{M}_{ij}$ is the average similarity between the vectors of the units with role i and the units with role j. In the hospital dataset, we show that the average similarity of the vectors of nodes with the same role is higher than that of nodes with different roles. This is particularly true for medics and nurses. However, this is not true for specific roles such as patients, who tend not to overlap much neither with other roles or with other patients. In a similar way, we notice that students from the same class in both high school and primary school datasets tend to be more similar than students from different classes. Moreover, some additional form of clustering emerges, probably due to the fact that classes that are more physically close have students that interact more, leading to some communities overlapping among those classes.

In Fig. 3.6b, we measure the average diversity of the community membership vectors for each role or class. We measure this property by considering the entropy of the community membership vector for each node, averaging nodes with the same role.

Let v_i be the community membership vector for node i, p_{ij} be the proportion of memberships of node i to community j (i.e., the number of distinct hyperlinks assigned to community j containing node i), and C be the set of communities node i participates in. The entropy H for node i is:

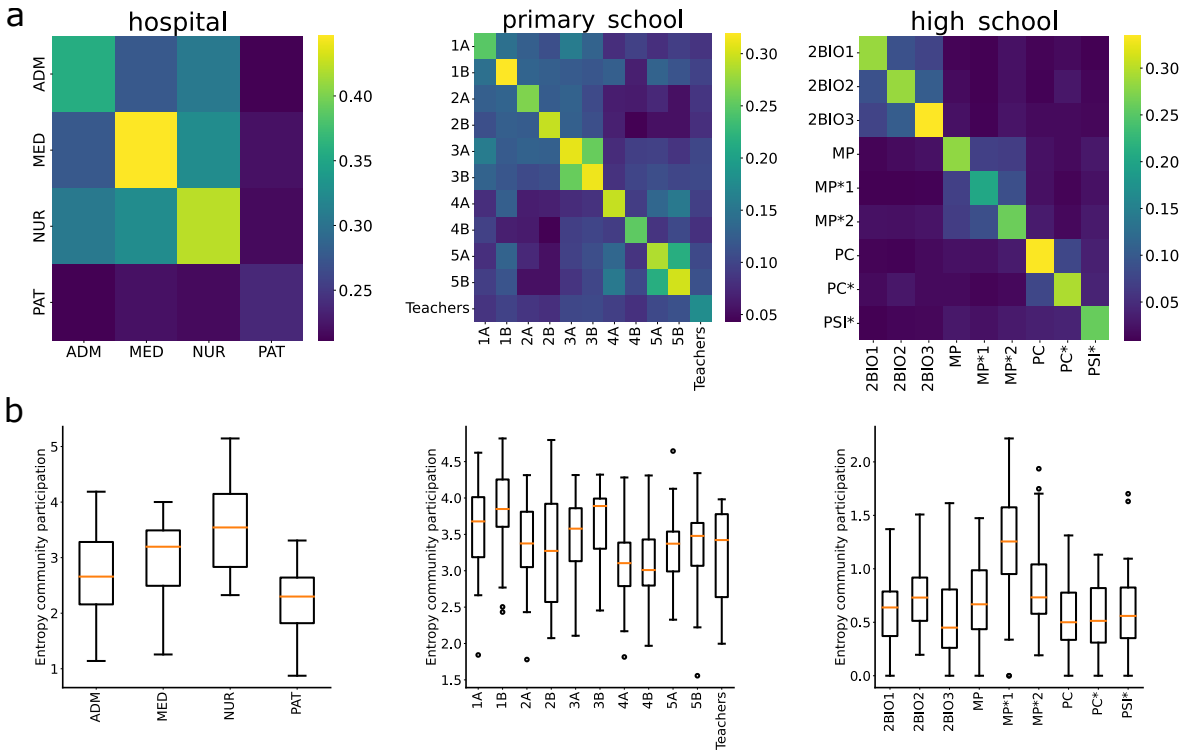

Figure 3.5: **Comparing overlapping communities and node metadata.** We select a threshold for cutting the dendrogram and extracting overlapping communities, and compare results with metadata from real-world hypergraphs (role or class). A binary community membership vector is used to identify whether a node participates in a certain community. **a)** We measure the pairwise similarity between the binary vectors (Jaccard similarity) and build the role-to-role similarity matrices by aggregating similarities of nodes based on their role. Nodes with similar roles or classes tend to share similar community memberships. However, patients in the hospital dataset have low overlapping memberships even with other patients. Moreover, clustering emerges among classes in the primary and high-school datasets, probably because their proximity leads to mixing interactions among different classes. **b)** We measure the diversity (entropy) of community membership vectors for each role or class averaging nodes with the same role. Certain roles, such as nurses, have a more diverse and pervasive overlap, while patients have less diversified interactions. In school datasets, some classes have more diverse community memberships, possibly due to physical constraints or participation in more activities.

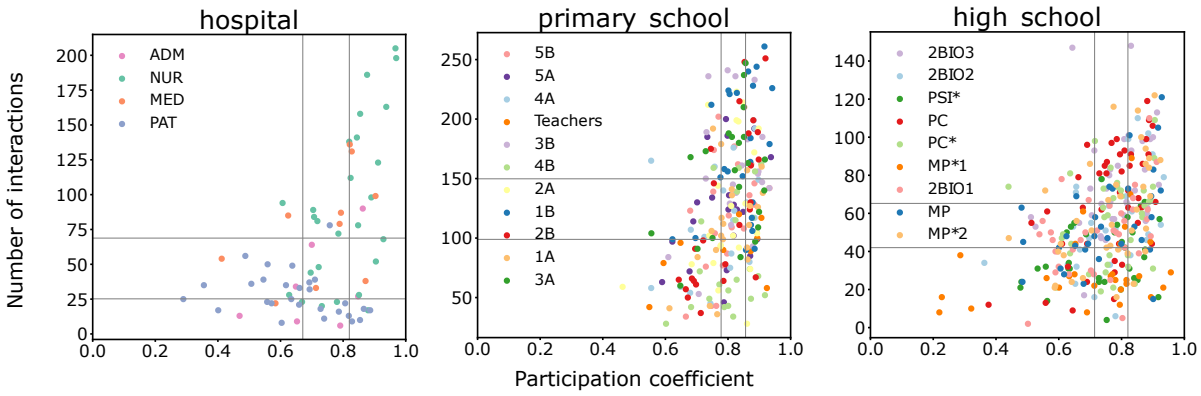

Figure 3.6: **Cartography of higher-order networks.** We provide a cartography of higher-order networks, where nodes are classified based both on their number of interactions (hyperdegree) and mixed-membership to hypergraph communities (participation coefficient). This classification yields nine structural roles (hub, non-hub, or peripheral on the y-axis; generalist, non-generalist, or specialist on the x-axis), where complementary information is provided by the two variables. We apply our method to three face-to-face higher-order social systems, showing how it can be used to capture metadata information. As an example, hospital patients tend to be peripherals but range from specialists to generalists. By contrast, in school data, each class has representatives of each structural role.

$$H(v_i) = -\sum_{j}^{C} p_{ij} \log(p_{ij}) \tag{3.2}$$

Let N_r be the number of nodes with role r; to average the entropy for nodes of the same role r we compute:

$$\bar{H}_r = \frac{1}{N_r} \sum_{i \in r} H(v_i) \tag{3.3}$$

We show that, indeed, different roles can have a more diverse set of memberships than others. For example, nurses share a more diverse and pervasive overlap than other roles, while patients are more strict and less overlapping. In the school datasets, there are classes that are more favoured than others in having multiple diverse community memberships, possibly due to physical constraints or participation in more activities.

3.4 Cartography of higher-order networks

In the previous sections, we have mentioned that nodes can be characterized in terms of their simultaneous participation in different communities. We have also highlighted that

the community memberships of nodes can range from very heterogeneous (i.e., membership vector with high entropy) to very homogeneous (i.e., membership vector with low entropy). Community participation behaviour of nodes can be linked with their hyperdegree to classify nodes into different structural roles.

We measure the hyperdegree of each node by considering the number of interactions they participate in. Community membership diversity is here computed by considering the participation coefficient of the membership vectors, paying homage to the seminal work by Guimera et al. on the cartography of complex networks [121] (analogous results are obtained by considering entropy). Such a score of diversity ranges from 0 (minimum diversity) to 1 (maximum diversity) and has also been used to characterize heterogeneity in different network structures such as multiplex networks [22]. In particular, having fixed a scale and computed overlapping communities, the participation coefficient of the membership vector of node i is defined as

$$P_i = \frac{C}{C-1}\left[\sum_{\alpha=1}^{C}\left(\frac{k_i^{[\alpha]}}{k_i}\right)^2\right] \tag{3.4}$$

where C is the number of communities, k_i is the hyperdegree of node i and $k_i^{[\alpha]}$ is the number of interactions of node i that are part of community α.

In Fig. 3.6, each node is represented as a point in the Cartesian plane with coordinates equal to its hyperdegree and participation coefficient. The nodes are classified into different regions by subdividing both the x-axis and the y-axis into the 33rd and 66th percentiles. On the y-axis, a node in the top third is classified as hub, in the middle third as non-hub and the bottom third as peripheral. On the x-axis, a node in the top third is classified as a generalist, in the middle third as a non-generalist and in the bottom third as a specialist. This classification gives a total number of nine structural roles, corresponding to different regions in the plane. Our results highlight that network nodes with different structural roles are indeed relevant for understanding real-world systems. Nodes with a similar hyperdegree can have a very different participation coefficient, and vice versa. Additionally, structural roles can be linked with metadata information. For example, in hospital data, we observe that people with the same jobs or status can have very different behaviours, e.g., patients tend to be peripherals but can range from specialists to generalists. Finally, in school data, in the same class, we have multiple representatives for each structural role.

3.5 Discussion

The analysis of network communities is widely recognized as a fundamental tool for understanding the interplay between the structure and dynamics of complex systems, finding applications in fields such as biology and social network analysis. Network communities are often defined as groups of nodes that are more densely connected to each other within the community than they are to nodes outside of the community. Recently, the framework of hypergraphs has established itself as a fundamental tool to model systems whose interactions are not limited to pairs of nodes but may involve an arbitrary number of them.

Using a dual approach to community detection, in this Chapter we have extended the traditional notion of link communities to hypergraphs, extracting clusters of highly related hyperedges. By defining a measure to determine the distance between two hyperedges and performing hierarchical clustering on top of the pairwise distance matrix of such hyperedges, we studied the dendrogram obtained as the output of such a process on a variety of real-world data. Hyperlink communities naturally highlight the hierarchical and multiscale structure of a higher-order network, at the same time revealing overlaps among node communities. Finally, we have introduced the notion of the cartography of higher-order networks, and classified nodes in different structural roles, i.e., a small number of system-independent roles that depend on the patterns of interactions of the nodes and their scale-specific overlapping community participation. We showed that with this classification we are able to capture information that the hyperdegree alone cannot provide.

Given the interest in link communities in pairwise networks, we believe that hyperlink community detection may serve as a relevant tool for analysing a variety of higher-order data, helping unveil structural patterns which cannot be explored with traditional community detection approaches. In this direction, interesting venues for future research include the evaluation of distance measures for hyperedges alternative to the Jaccard distance used here, and the development of methods to scale the computation of hyperlink distances in very large real-world hypergraphs.

4 | The microscale organization of directed hypergraphs

So far, we focused on and proposed tools to study undirected hypergraphs, failing to capture a potentially interesting feature of many real-world interactions: their directional nature. For example, in a metabolic reaction, a set of reactants transforms into a set of products [122]. Similarly, in a Bitcoin transaction, multiple source wallets may transfer funds simultaneously to multiple target wallets [31]. Directed hypergraphs enhance modeling by distinguishing between source and target sets in each hyperedge [123] to accurately encode directionality. Tools to study directed hypergraphs are largely underdeveloped, with notable exceptions in areas such as null models [124], synchronization [125], overlapping patterns between two hyperedges of limited size [126], and some early proposals to define reciprocity [127, 128].

In this Chapter, we introduce measures and tools to characterize the microscale organization of real-world directed hypergraphs. We decompose interactions into four types (one-to-one, one-to-many, many-to-one and many-to-many) and analyze their prevalence in empirical data to highlight differences in higher-order connectivity across domains. We investigate the overlap among sources and targets to reveal recurring sets of co-sending and co-receiving nodes. We also propose definitions for exact, strong, and weak higher-order reciprocity to capture patterns of bi-directionality and extend motif analysis to include directed higher-order interactions.

4.1 Directed hypergraphs

Traditional graph models reduce directed group interactions into a collection of pairwise links, often leading to a loss of important structural information about group organization and dynamics. For instance, reducing a many-to-many interaction such as SOURCE $= \{A, B\}$ and TARGET $= \{D, E\}$ to a set of pairwise directed links ($A \to D$, $A \to E$, $B \to D$ and $B \to D$) fails to capture the collective nature of the interaction, where pairs of nodes are jointly involved in the source and target sets. Directed hypergraphs preserve

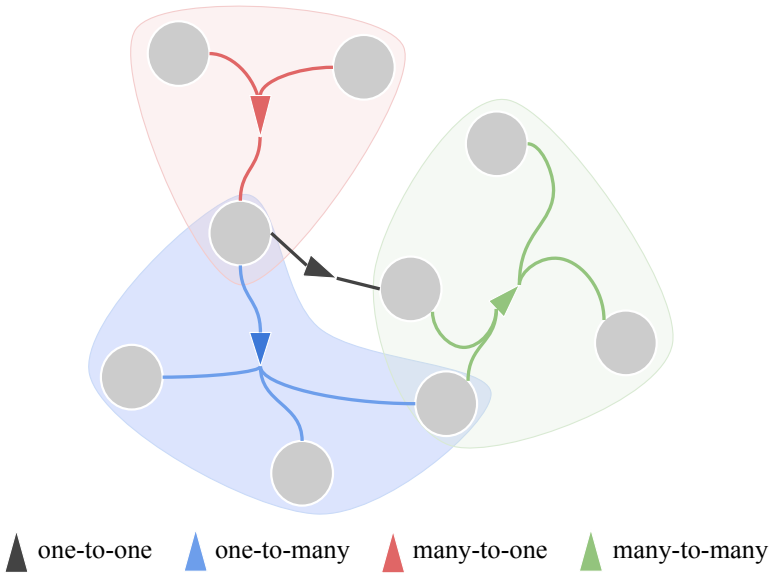

Figure 4.1: **Schematic of a directed hypergraph.** Each interaction encodes a source set of units acting towards a target set of units. We distinguish four types of directed higher-order interactions: one-to-one (black), one-to-many (blue), many-to-one (red), and many-to-many (green).

group-based structure, allowing for a more faithful representation of complex interactions. In such a mathematical framework, hyperedge direction is encoded by distinguishing between source and target node sets, which are non-empty and disjoint. In particular, we distinguish four fundamental patterns of interactions encoded as directed hyperedges: one-to-one, where a single source node connects to a single target; one-to-many, where one source affects multiple targets; many-to-one, where multiple sources act on a single target; and many-to-many, the most general case, where multiple sources act on multiple targets. In Figure 5.1, we show an example of a directed hypergraph, highlighting all distinct hyperedge patterns.

To study the microscale organization of real-world systems with directed group interactions, we collected datasets from multiple domains and mapped them into directed hypergraphs. The datasets [128] include QNA (nodes are users and forum posts are hyperedges), E-MAIL (nodes are users and emails are hyperedges), BITCOIN (nodes are accounts and financial transactions are hyperedges), METABOLIC (nodes are genes and metabolic reactions are hyperedges) and CITATION (nodes are authors and hyperedges are paper citations). Detailed descriptions and summary statistics of each dataset are available in Appendix A.2.

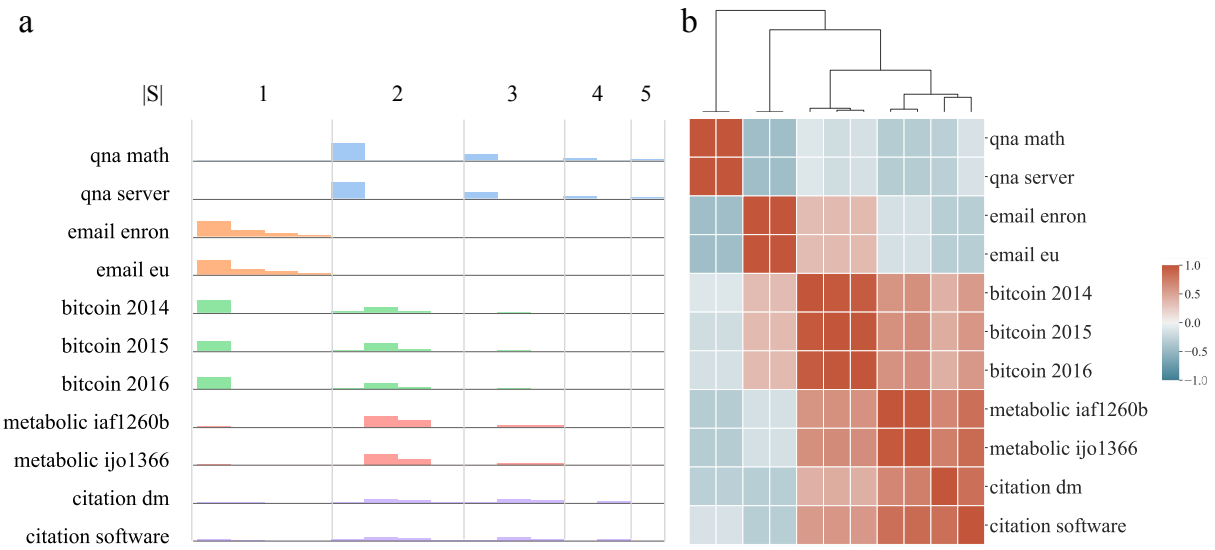

Figure 4.2: **Hyperedge signature of directed hypergraphs.** a) We characterize each system with a signature hyperedge vector, encoding the abundance of a certain pattern of directed hyperedge. Vectors are normalized. $|S|$ indicates the cardinality of the respective part of the vector. Statistics are computed for hyperedges of cardinality at most 6. Systems within the same domain share the same color. b) Dendrogram resulting from agglomerative clustering applied to the correlation matrix of hyperedge signature vectors for each dataset. Correlation values are color-coded, with high positive correlations in red and high negative correlations in blue.

4.2 Patterns of directed hyperedges

We characterize directed hypergraphs across domains by investigating the diversity in their patterns of directed hyperedges. For each dataset, we construct a *hyperedge signature vector* $\mathbf{v}$, where each element represents the count of hyperedges with a specific combination of source set size s and target set size t in the hypergraph. The vector $\mathbf{v}$ captures the distribution of hyperedges based on the sizes of their source and target sets. Formally, we define the vector $\mathbf{v}$ as follows:

$$\mathbf{v} = (v_{1,2}, \ldots, v_{1,K-1}, v_{2,1}, \ldots, v_{2,K-2}, \ldots, v_{K-1,1})$$

where K represents the maximum hyperedge size considered, and each $v_{h,t}$ counts the number of hyperedges with a specific source size h and target size t. Such vectors provide a fingerprint for systems based on their higher-order connectivity patterns at the microscale. Figure 4.2a shows the hyperedge signature vectors for each dataset, considering interactions up to size 6. To emphasize the role of higher-order interactions in the

analysis, we do not consider one-to-one interactions. In the E-MAIL data we find abundance only in entries corresponding to one-to-many interactions, reflecting the typical structure of email communications. Similarly, in the QNA, many-to-one interactions are prevalent, as these systems involve multiple individuals responding to a question by a single user. In contrast, METABOLIC and CITATION datasets show high abundances in many-to-many relationships across a variety of source and target set sizes. Finally, BITCOIN dataset exhibits more varied behavior, with abundant entries for both one-to-many and many-to-many interactions, indicating different interaction types in the network.

To further explore structural diversity across different domains, we compute pairwise correlations between hyperedge signature vectors (Pearson coefficient) and apply hierarchical agglomerative clustering on their correlation matrix. A correlation value close to 1 indicates similar hyperedge structures, 0 suggests no relationship, and -1 indicates the structures are inversely related. The clustering procedure applied to the systems' correlation matrix results in a dendrogram that visually represents their hierarchical relationships, highlighting the presence of clusters of directed hypergraphs that share similar connectivity patterns. In Figure 4.2b, we show the correlation matrix and the clustering dendrogram. By examining the correlation matrix, we observe a strong correlation within systems from the same domain, indicating highly similar abundance in hyperedge structures. In contrast, systems from different domains exhibit varying degrees of correlation. Specifically, E-MAIL and QNA datasets are inversely correlated, as they display non-overlapping and complementary connectivity patterns: E-MAIL is characterized by one-to-many interactions, whereas QNA primarily involves many-to-one relationships. The METABOLIC and CITATION datasets, which feature many-to-many interactions, are positively correlated and form a distinct cluster. Interestingly, the BITCOIN datasets also display positive correlations with the METABOLIC and CITATION cluster due to a high presence of many-to-many interaction patterns. However, they also exhibit a weaker positive correlation with the E-MAIL datasets, reflecting the presence of one-to-many interactions in BITCOIN.

4.3 Source and target sets overlap

Nodes are frequently involved in multiple hyperedges, either as part of the source or the target set. The degree to which hyperedges share elements provides valuable insights into redundancy, hierarchical structures, and information flow within the system. To quantify this, we introduce the concept of *excess overlap*, which measures for each node the overlap among the source and target sets in which it participates compared to what would be

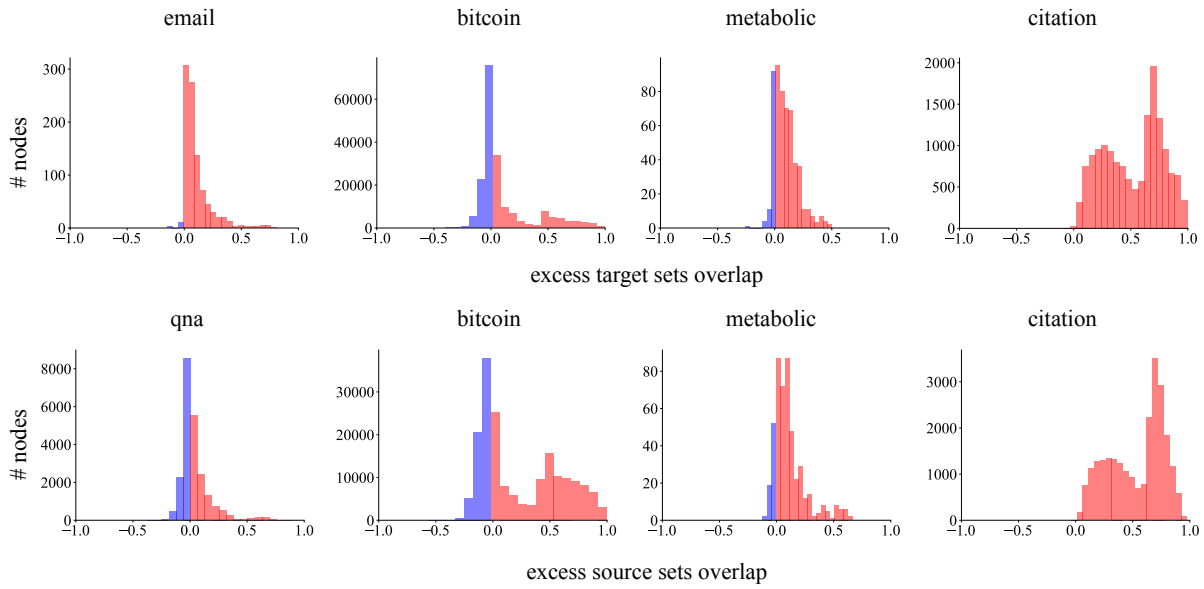

Figure 4.3: **Excess of overlap across domains.** The top row shows the distribution of target sets overlap, while the bottom row depicts the source sets overlap distribution. Red indicates positive overlap values, meaning interactions are more structurally redundant. Blue indicates negative overlap values, suggesting that interactions are less repetitive, implying a tendency for diverse co-sender and co-receiver combinations compared to random expectation.

expected under a null model. For each node, we quantify its hyperedge overlap separately for its participation in target sets (in-hyperedges) and source sets (out-hyperedges). For in-hyperedge overlap, consider all hyperedges e for which the node is in the target set $t(e)$. Let $\mathcal{E}_{\text{in}}$ be the collection of these hyperedges. We define the in-hyperedge overlap as

$$O_{\text{in}} = \frac{\sum_{e \in \mathcal{E}_{\text{in}}} |t(e)|}{|\mathcal{E}_{\text{in}}| \left| \bigcup_{e \in \mathcal{E}_{\text{in}}} t(e) \right|},$$

where $|t(e)|$ denotes the number of nodes in the target set of hyperedge e. Similarly, for out-hyperedge overlap, we consider only hyperedges e in which the node is a source (i.e., belongs to $s(e)$). Let $\mathcal{E}_{\text{out}}$ denote this collection, and define

$$O_{\text{out}} = \frac{\sum_{e \in \mathcal{E}_{\text{out}}} |s(e)|}{|\mathcal{E}_{\text{out}}| \left| \bigcup_{e \in \mathcal{E}_{\text{out}}} s(e) \right|},$$

with $|s(e)|$ being the size of the source set of hyperedge e. These metrics yield a value of 1 when all corresponding hyperedges share an identical set of nodes (i.e., maximal overlap), and decrease as the sets become more diverse. To assess statistical significance,

we compare the observed overlaps to those computed on an ensemble of randomized networks that preserve key structural properties (e.g., node degrees and hyperedge sizes). For each node, the excess overlap ΔO is defined as the difference between its observed overlap and the average overlap derived from these null models:

$$\Delta O = O^{\text{obs}} - \langle O^{\text{rand}} \rangle.$$

We compute this for both O_{in} and O_{out}. When the observed overlap exceeds the expected overlap from a random model (excess overlap > 0), it indicates that nodes tend to participate in structurally redundant interactions, where the same groups of nodes frequently co-occur in source or target sets. In contrast, when the observed overlap is lower than expected (excess overlap < 0), it suggests that interactions are more diverse, with hyperedges being more distinct and less likely to share members.

In Fig. 4.3, we show the distribution of the excess of overlap for source and target sets across domains. All datasets generally exhibit a high degree of overlap, indicating that recurrent behaviors are a shared feature. The CITATION dataset displays significant overlap for both source and target sets, showing that (i) an author tends to preferentially work with known collaborators, and (ii) an author tends to be cited repeatedly alongside similar sets of authors. The METABOLIC dataset follows a similar trend, displaying significant excess overlap for both source and target sets. This suggests that metabolic reactions tend to involve recurring sets of substrates and products and highlights the modular nature of metabolic networks. The BITCOIN dataset exhibits generally high excess overlap for both source and target sets. However, the presence of nodes with overlap values lower than zero implies that certain participants in the network engage in interactions that introduce more novelty rather than reinforce existing hyperedges. In the E-MAIL dataset, the excess overlap can be computed only for the target sets, as the source sets always have cardinality 1. The overlap is significantly larger than random, underscoring the hierarchical and broadcast-like nature of email communication. QNA data show a lower excess overlap compared with the other systems, indicating that forum respondents are less likely to engage repeatedly with the same set of co-responders. Since the target sets in this dataset always have cardinality 1, the excess overlap can be computed only for source sets.

4.4 Higher-order reciprocity

Reciprocity is a fundamental property of systems with directed interactions, including social networks [129]. It traditionally refers to the tendency of the system's units to mu-

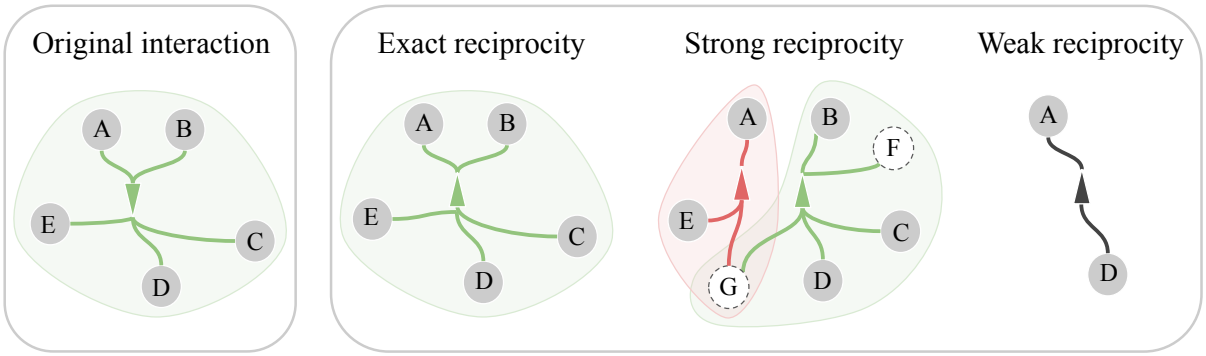

Figure 4.4: **Reciprocity measures for directed hypergraphs.** For exact reciprocity, the direction of a hyperedge is fully reversed by a single hyperedge in which the source and target sets are swapped. In strong reciprocity, multiple hyperedges collectively reverse the interaction, with the source and target sets being fully reciprocated through a combination of interactions. In weak reciprocity, at least one node from the target set reciprocates an interaction with one node from the source set.

tually exchange information. In directed graphs, reciprocity is defined as $r = \frac{\overleftrightarrow{L}}{L}$, i.e., the ratio of the number of bidirectional links ($\overleftrightarrow{L}$) to the total number of links (L). This measure has been widely used to describe real-world directed networks [19, 130]. Recognizing its broad importance, recent works have extended reciprocity to hypergraphs, accounting for the complexity of having multiple nodes in both the source and target sets of hyperedges. Among the recent approaches for hypergraph reciprocity, one method decomposes hyperedges into pairwise links [127], losing information about group interactions. An alternative approach defines a more complex measure that diverges from the traditional binary definition of reciprocity at the level of single links [128]. While this approach can capture different nuances, it is computationally expensive and less straightforward to interpret, as it provides a continuous value instead of a simple yes-or-no answer to whether an interaction is reciprocated.

Here, we introduce three simple and computationally efficient measures for higher-order reciprocity in directed hypergraphs, capturing different aspects of mutual interactions:

- **Exact reciprocity** occurs when an interaction represented by a hyperedge with a source set h and a target set t is precisely mirrored by another interaction with the source and target sets reversed. Formally, two hyperedges $e_1 = (h_1, t_1)$ and $e_2 = (h_2, t_2)$ are exactly reciprocated if and only if $h_1 = t_2$ and $t_1 = h_2$. This is the strictest form of reciprocity.
- **Strong reciprocity** relaxes the previous requirement and allows source and target

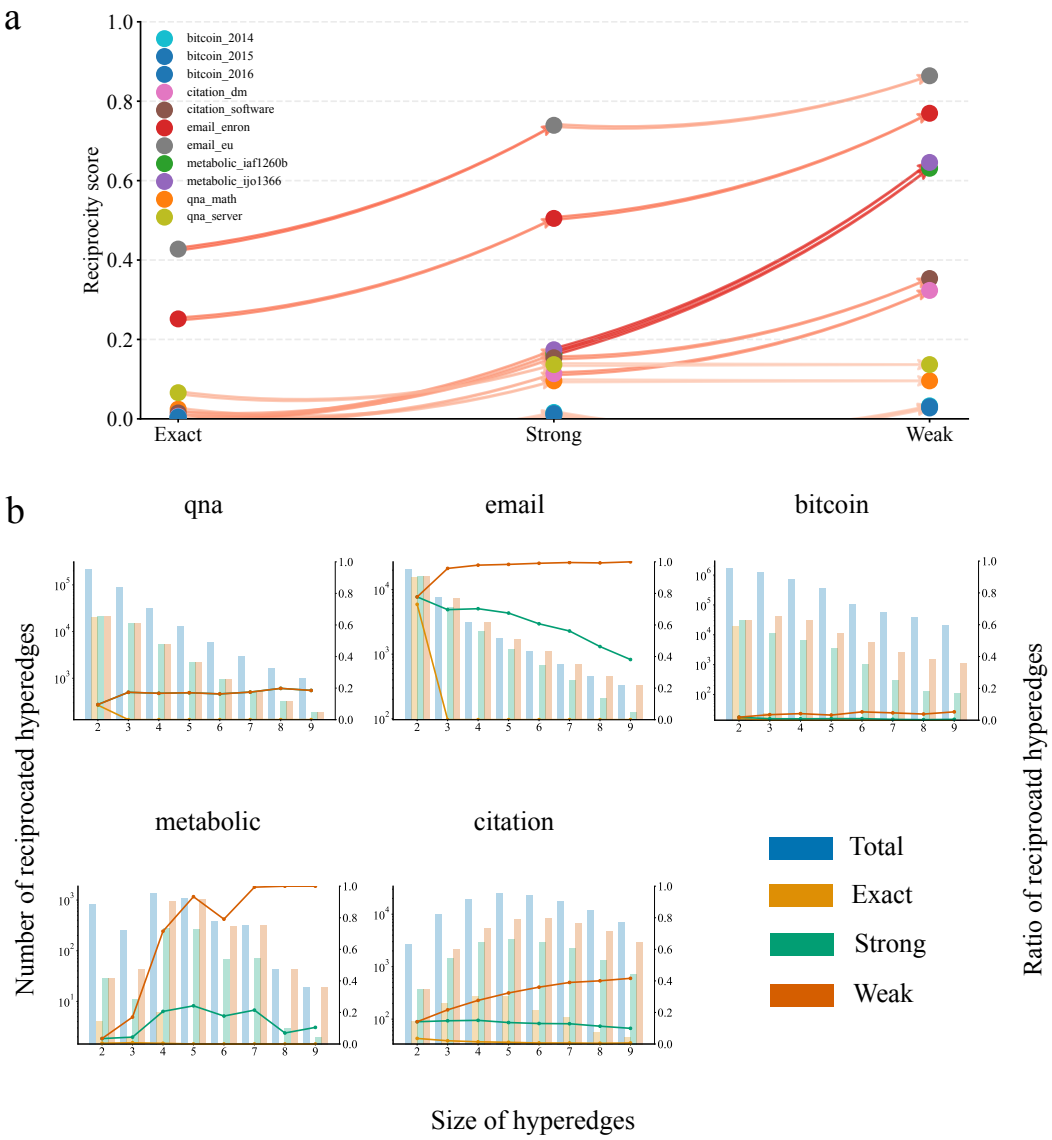

Figure 4.5: **Higher-order reciprocity in real-world hypergraphs.** a) Ratio of reciprocated hyperedges across datasets and reciprocity definitions. Each column corresponds to a distinct notion of higher-order reciprocity, thereby inducing a ranking of the datasets based on their reciprocity scores. Datasets are represented in unique colors. Red arrows link the same dataset across different definitions, with darker arrows indicating larger shifts in scores and lighter arrows representing smaller changes. b) Number of reciprocated hyperedges for each different notion of reciprocity. Statistics are disaggregated by hyperedge size. In blue, total hyperedges; in yellow, exactly reciprocated hyperedges; in green, strongly reciprocated hyperedges; and in orange, weakly reciprocated hyperedges. We use lines of the same colors to depict the ratio of reciprocated hyperedges with respect to total hyperedges for each notion of reciprocity. To simplify the plot, we grouped higher-order reciprocity of systems from the same domain.

sets to be reversed through a combination of hyperedges, instead of requiring a direct reversal with a single opposite one. Formally, a hyperedge $e = (h, t)$ is *strongly reciprocated* if there exists a set of hyperedges $\{e_1, e_2, \ldots, e_k\}$ such that the union of the target sets of $e_1, \ldots, e_k$ is a superset of the source set h, and the union of the source sets of $e_1, \ldots, e_k$ is a superset of the target set t.

- **Weak reciprocity** represents the most relaxed form of reciprocity and requires only that at least one node from the target set of a hyperedge appears in the source set of another, and vice versa. Formally, a hyperedge $e = (h, t)$ is *weakly reciprocated* if there exists another hyperedge $e' = (h', t')$ such that $h \cap t' \neq \emptyset$ and $t \cap h' \neq \emptyset$.

We summarize our definitions of reciprocity for directed hypergraphs in Figure 4.4. More information about the algorithmic aspects of such measures is available in Appendix C.2.

After introducing these definitions, a natural first question is which systems exhibit the highest and lowest levels of reciprocity and how the ranking of systems based on reciprocity changes across different definitions. We address this in Figure 4.5a, which shows the ratio of reciprocated hyperedges (reciprocity score) for each system across varying notions of reciprocity. The reciprocity score for each definition induces a ranking of the systems, allowing us to observe which systems exhibit stronger tendencies toward mutual exchange of information. By definition, the score tends to increase for each system as we move from stricter definitions of reciprocity (exact) to more relaxed ones (weak). We observe that systems from the same domain tend to show similar levels of reciprocity across definitions, indicating that functional similarities within domains may drive comparable reciprocity patterns. E-MAIL datasets exhibit the highest levels of reciprocity, while BITCOIN datasets consistently show the lowest. Interestingly, while the ranking of systems remains largely stable with varying definitions, the relative distances between the datasets change. For instance, exact reciprocity mostly characterizes E-MAIL datasets, which are positioned far from the other datasets, clustering distinctly at the top of the scale. Strong reciprocity induces three clear clusters of datasets based on their scores: E-MAIL datasets rank the highest by a large margin, while BITCOIN datasets occupy the very low end. In the case of weak reciprocity, the datasets begin to separate along domain lines, spanning the entire spectrum of reciprocity scores. Notably, we observe a reduction in the distance between E-MAIL and datasets from metabolic and citation domains, suggesting a convergence in reciprocity levels as the definition becomes more relaxed. Overall, these patterns highlight how the choice of measure can influence the perceived level of reciprocity within different systems. By analyzing how the score evolves across definitions, we gain a more precise understanding of the extent of mutual exchange within each system, from the high reciprocity observed in E-MAIL datasets, where high mutual exchange is clear, to the

lower reciprocity in BITCOIN datasets, where reciprocal connections are minimal across all definitions, and to the metabolic datasets, which emerge with high reciprocity under weaker definitions.

A related question is how the size of hyperedges influences the levels of reciprocity. We explore this in Figure 4.5b, which presents statistics disaggregated by interaction size and shows both the number and ratio of reciprocated hyperedges across systems and reciprocity definitions. Notably, exact reciprocity in the E-MAIL and QNA datasets is only possible for dyadic links because, for example, a single e-mail cannot have multiple senders simultaneously. In general, exact reciprocity is prevalent in the E-MAIL datasets but is concentrated only in dyadic interactions. We also observe exact reciprocity in the QNA datasets, but again, only for dyadic interactions. In the CITATION dataset, exact reciprocity is present across interactions of all sizes, albeit to a smaller degree. Overall, exact reciprocity tends to decrease as hyperedge size increases. Strongly reciprocated hyperedges are common across all systems, but their ratio declines with hyperedge size at different domain-dependent rates. Weak reciprocity, however, is widespread across systems and tends to increase with hyperedge size, indicating that larger interactions tend to exhibit more mutual exchange when reciprocity is defined more loosely. These findings suggest that weaker notions of reciprocity are valuable in providing insights into the overall reciprocity of systems with larger interactions.

4.5 Motif analysis in directed hypergraphs

We recall that motif analysis involves counting the frequency of patterns of interactions in connected subgraphs of a given number of nodes. This framework was first introduced by Milo et al. [60] to extract the fundamental functional units of complex systems [61]. In Chapter 2, we have extended motif analysis to hypergraphs to capture patterns of interactions with arbitrary size. Here, we extend such analysis to consider also the direction of the hyperedges involved in the patterns.

First, it is interesting to study the combinatorics of the patterns of directed subhypergraphs. There is no simple closed-form formula for counting the number of possible directed higher-order motifs as a function of their order n, i.e., the number of nodes in the patterns. We can estimate the number of non-isomorphic connected directed hypergraphs in a way similar to [1]. Given a set of n nodes, the number of possible directed hyperedges is $3^n - 2 \cdot \sum_{k=1}^{n} \binom{n}{k} - 1 = 3^n - 2 \cdot 2^n + 1$. This expression counts the ways to partition the n nodes into three disjoint sets: source, target and empty set. We subtract the invalid combinations with empty source or target sets. Given n nodes, we ensure connectivity by

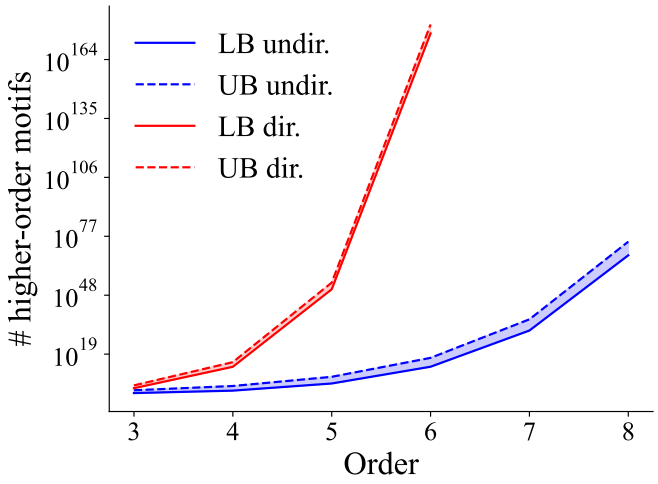

Figure 4.6: **Combinatorics of directed higher-order motifs.** Upper (dashed lines) and lower (solid lines) bounds on the number of higher-order motifs as a function of their order. Blue lines refer to undirected motifs on hypergraphs, red lines refer to the directed case.

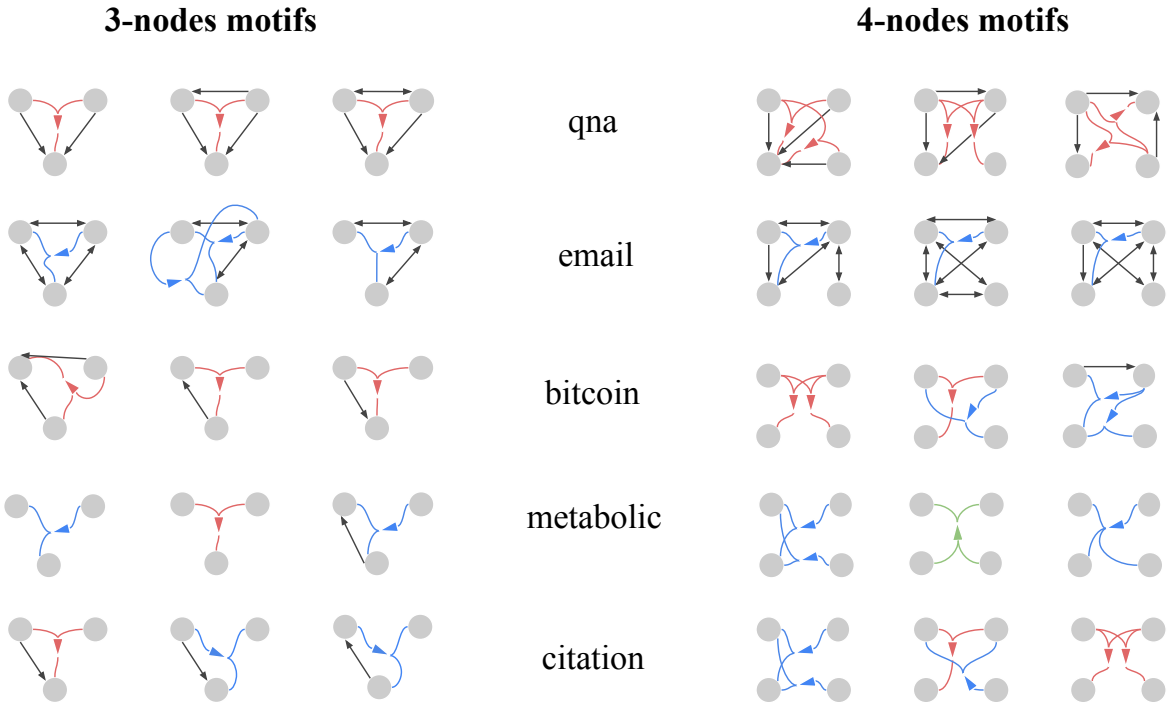

Figure 4.7: **Directed higher-order motifs in real-world hypergraphs.** The three most representative directed higher-order motifs of orders three and four from each system. The color of a group interaction encodes its type: one-to-one (black), one-to-many (blue), many-to-one (red), and many-to-many (green). We group statistics of systems within the same domain.

selecting a chain of $n-1$ hyperedges and including them in the hypergraph, leaving us with $3^n - 2 \cdot 2^n - n + 2$ remaining possible hyperedges. For each remaining hyperedge, we decide whether to include it or not, resulting in $2^{3^n - 2 \cdot 2^n - n + 2}$ total hypergraphs. Since we are interested in non-isomorphic hypergraphs, we divide this number by $n!$, the number of ways to label the vertices, providing the lower bound $\frac{2^{3^n - 2 \cdot 2^n - n + 2}}{n!}$. If we ignore the constraints of non-isomorphism and connectivity, we count the number of possible labeled hypergraphs. Since each of the $3^n - 2 \cdot 2^n + 1$ possible hyperedges can either be included or excluded, the total number of labeled hypergraphs is at most $2^{3^n - 2 \cdot 2^n + 1}$. Figure 4.6 shows the upper and lower bounds on the growth of possible sub-hypergraph patterns as a function of the number of nodes (order), for both the undirected and directed cases. The estimated number of patterns grows super-exponentially, even in the undirected case. In the directed case, the growth is even faster due to the need to consider all possible subdivisions into source and target sets.

We propose an exact algorithm to count the frequency of all connected sub-hypergraph patterns, following ideas from Chapter 2. An extensive presentation of the algorithms for motif analysis in directed hypergraphs, including pseudocode and performance evaluation, is available in Appendix C.3. To distinguish meaningful, non-random interaction patterns from those that may occur by chance, we use a configuration model as a null model to evaluate the statistical significance of the interaction patterns after computing their frequency in our directed hypergraphs. The configuration model generates randomized versions of the original hypergraph while preserving key properties, such as the in-degree and out-degree sequences, as well as the source and target sizes of the hyperedges [124]. By comparing the observed frequencies with those found in the randomized networks, we can identify significantly over-represented motifs. In particular, each motif i is associated with the abundance score Δ_i relative to random networks proposed in [61]. This abundance score is defined as follows:

$$\Delta_i = \frac{N\text{real}_i - \langle N\text{rand}_i \rangle}{N\text{real}_i + \langle N\text{rand}_i \rangle + \epsilon} \tag{4.1}$$

where $N\text{real}_i$ is the frequency of motif i in the empirical hypergraph and $\langle N\text{rand}_i \rangle$ is the average frequency of motif i across multiple realizations of the configuration model. Following [1, 61], we set $\epsilon = 4$. We sample 10 times from the configuration model. Given the intractability of the problem for large sub-hypergraphs, we limit our study of empirical data to patterns involving three and four nodes. Moreover, we focus on patterns that include at least one group interaction.

In Figure 4.7, we show the most over-represented patterns of directed higher-order inter-

actions with three and four nodes across different domains. Each domain reveals distinct motifs, characterized by different directed hyperedge types, sizes, densities and patterns of reciprocity. In terms of hyperedges types, E-MAIL and QNA involve abundant patterns with only many-to-one and one-to-many interactions. Other datasets display more diverse patterns, including combinations of one-to-many, many-to-one, and many-to-many interactions (the latter is possible only in motifs with four nodes). Traditional one-to-one interactions are commonly part of abundant patterns in all datasets. The number of interactions in abundant sub-hypergraphs is small in the BITCOIN, METABOLIC and CITATION domains, often involving just one or two hyperedges. In contrast, the E-MAIL and QNA domains tend to be richer in interactions. This observation is reversed when considering the average size of interactions. The relation between the number and the average size of interactions aligns with previous studies on undirected higher-order motifs [1]. A common pattern in many datasets is the coexistence of group interactions alongside lower-order interactions within the same set of nodes. These interactions seem to play a role in increasing the overall reciprocity of the patterns, suggesting the existence of a feedback mechanism. This is particularly evident in E-MAIL data. In addition to reciprocity, the direction of lower-order interactions in abundant patterns suggests a reinforcing mechanism where subsets of source and target nodes interact at multiple interaction sizes. These observations are closely connected with the insights discussed in the previous sections about frequent co-senders and co-receivers nodes and higher-order reciprocity.

4.6 Discussion

Directed hypergraphs enhance our modelling abilities by accounting for directionality in group interactions, distinguishing between source and target sets for each hyperedge. This versatile framework can accurately model a range of diverse real-world systems and interactions, including financial transactions, email exchanges, and metabolic reactions.

In this Chapter, we proposed new measures and tools to analyze the structural organization of directed hypergraphs at their microscale. First, we analyzed hyperedge signature vectors to identify the abundance of each hyperedge structure across datasets and identified classes of systems sharing similar higher-order connectivity patterns. Second, we analyzed the excess overlap among source and target sets for each node in each system. The resulting distributions suggest that different domains may follow distinct organizational principles, ranging from redundant to more diverse interaction patterns. Then, we introduced three distinct types of higher-order reciprocity measures: exact, strong, and

weak reciprocity. Each definition offers a different perspective on how group interactions can be reciprocated, ranging from strict to more relaxed forms of reciprocal influence, and can be computed efficiently, making it suitable also for the analysis of very large systems. We showed that all systems exhibit reciprocity in broad terms, though different domains are associated with specific patterns and sensitivity to specific reciprocity measures. Lastly, we extended the notion of motifs to directed hypergraphs, capturing recurring patterns of directed interactions. Motif analysis revealed frequent microscale structures and highlighted common organizational principles playing a role in the function and behavior of systems, such as the existence of reinforcing or feedback mechanisms among dyadic and non-dyadic interactions in groups.

Taken together, by considering the nuances related to the directionality of interactions in directed hypergraphs, we provide a framework to understand higher-order connectivity in directed complex systems, opening up a wide range of potential applications in diverse fields such as social network analysis, biology, and finance. For instance, the study of multi-party financial transactions as directed higher-order structures may capture more complex patterns of fraudulent activity than traditional graph-based models [131]. Similarly, directed hypergraphs may enhance the accuracy of existing frameworks in identifying and predicting important genes based on genomic expression relations [132]. As scalability is a pressing issue in hypergraph algorithms, future work may explore advanced techniques for detecting motifs in large-scale directed hypergraphs, including sampling methods [2], to expand our analysis beyond patterns of four nodes. Another interesting venue for further studies is related to the study of reciprocity in weighted or time-evolving hypergraphs, where interactions are associated with different intensities or specific moments in time.

5 | Multiplex measures for higher-order networks

Not all interactions in complex systems are alike: they may differ in nature, type, and scope. This observation led researchers to introduce the concept of multilayer and multiplex networks [133, 134], where links are encoded into different interaction layers, each representing a distinct type of relationship [21, 22]. Multilayer and multiplex networks can successfully describe systems such as trade networks [135], transportations networks [136], collaboration networks [137], and the brain [138]. Multiplex hypergraphs, where layers encoding hyperedges of different type, could offer a robust tool for describing complex systems that involve group interactions of varying types. Despite significant potential, multiplex hypergraphs remain relatively unexplored, and a general set of tools for their analysis is still missing.

In this Chapter, we introduce measures to characterize multiplex networks with higher-order interactions across different scales, from node activity patterns to mesoscale organization. We validate these measures on three real-world systems revealing distinct patterns of group interactions across different contexts and domains.

5.1 Multiplex hypergraphs

Multiplex hypergraphs model systems where interactions among units (i) may belong to multiple types and (ii) are not necessarily dyadic, i.e. they may involve more than two units. A *multiplex hypergraph* $\mathbf{H}$ is defined as:

$$\mathbf{H} = \{H_1(V, E_1), \ldots, H_M(V, E_M)\}$$

where each *layer* α is a hypergraph $H_\alpha(V, E_\alpha)$. Each hypergraph $H_\alpha(V, E_\alpha)$ share the same set of entities V. $E_\alpha \subseteq \mathcal{P}(V)$ is the set of all interactions of a specific type α. Moreover, we require $|e| \geq 2$ for all $e \in E_\alpha$ for any α. In other words, each layer in our framework shares the same set of nodes and represents a distinct set of interactions of

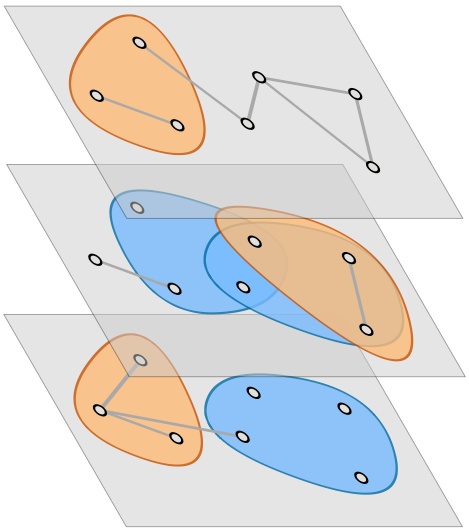

Figure 5.1: Multiplex hypergraphs represent systems of units that display interactions of different orders and different types. Each type of interaction is encoded into a single layer of the hypergraph. All the layers share the same set of nodes.

the same nature. In Fig. 5.1, we show a simple multiplex hypergraph with 7 nodes and hyperedges spread across 3 layers.

We remark that our proposed framework is different from a multiplex representation of higher-order interactions where layers are defined by interactions of different order [51, 139].

Here, we introduce a set of general tools to investigate multiplexity across different system scales in networks with higher-order interactions. We validate our measures and discuss relevant findings for three distinct real-world datasets:

- APS (Co-authorship network), where nodes are authors, and hyperedges represent groups of authors who have co-authored a paper. Each layer collects papers from the same subfield of physics, identified by a PACS code (Physics and Astronomy Classification Scheme) [87].
- IMDB (Co-starring network), where nodes represent actors, and hyperedges represent the cast of a specific movie. Each layer corresponds to a movie genre.
- HIGH SCHOOL (Social network), where nodes are students, and hyperedges represent groups of students interacting face-to-face, with each layer grouping interactions from the same day of the week [14].

Detailed and layer-disaggregated statistics about the datasets are provided in Table 5.1.

APS				IMDB				HIGH SCHOOL			
Layer	N	E	d	Layer	N	E	d	Layer	N	E	d
AMPhys	30375	12562	3.7	Anim	5545	864	9.5	Mo	312	2655	1.3
CM1	63919	27241	3.2	Comedy	69303	9111	13.6	Tu	310	3002	1.2
CM2	103075	58075	4.0	Doc	13357	2007	7.4	We	303	2543	1.2
EMag	57056	30908	2.8	Drama	103163	15384	12.6	Th	295	2529	1.2
EPart	62997	26703	62.2	Family	12968	1274	12.6	Fr	299	2339	1.2
GAA	41535	10670	9.6	Fantasy	14793	1136	15.0				
GasPhy	16182	5120	5.3	Horror	28254	2964	11.6				
Gen	69074	35940	3.2	Thriller	44188	4739	13.8				
IntPhy	48136	17382	3.0								
NPhy	50142	20672	16.7								
H	315421	219769	12.2	**H**	195377	37465	12.6	**H**	327	7818	1.3

Table 5.1: Statistics about three real-world multiplex hypergraphs. For each layer, we report the number of active nodes N, the number of hyperedges E and their average order $\bar{d}$. In the case of multiple PACS codes or genres associated with a paper or movie, only one code or genre is randomly selected. **H** is the layer-aggregated hypergraph.

5.2 Node properties

We begin by investigating multiplex properties at the node level. The first basic measure we consider is *node activity* [140]. A node i is *active* at layer α if i participates in at least one interaction at layer α. Fig. 5.2 shows statistics on nodes' simultaneous activity across multiple layers. Specifically, the y-axis plots the proportion of nodes (from the total node count) active in at least x layers. By definition, these curves exhibit a decreasing trend, with variations in the negative slopes reflecting the datasets' diversity. While it is uncommon for scientists and actors to be active across more than 2 or 3 layers, students tend to be active in all layers. In fact, the inactivity of a student in a specific layer implies their absence from school on that day.

So far, we have grouped all interactions together, regardless of their order. To obtain more detailed insights about higher-order interactions, we can examine node activity for each specific interaction order d. To this scope, we introduce a list $\mathbf{A}$ of *node activity matrices*, one for each node i:

$$\mathbf{A}_i = \{a_{\alpha d}\} = \begin{cases} 1 & \text{if } i \text{ is active at order } d \text{ in } \alpha \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, one can define activity from a layer perspective and consider a list $\mathbf{B}$ of *layer*

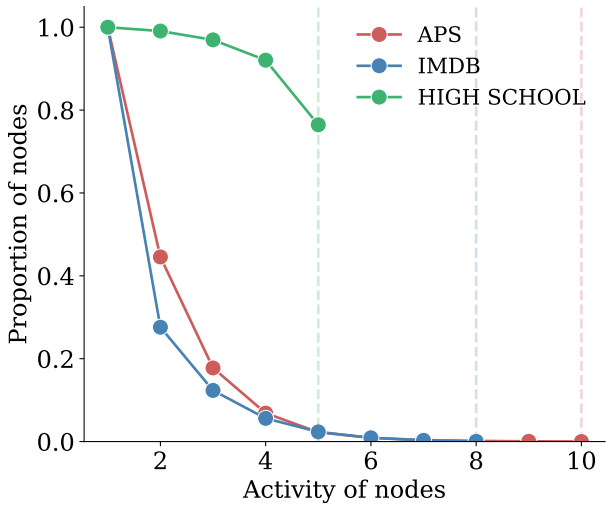

Figure 5.2: Proportion of nodes active in at least x layers across three different datasets. Colored dashed lines indicate the number of layers in each respective dataset.

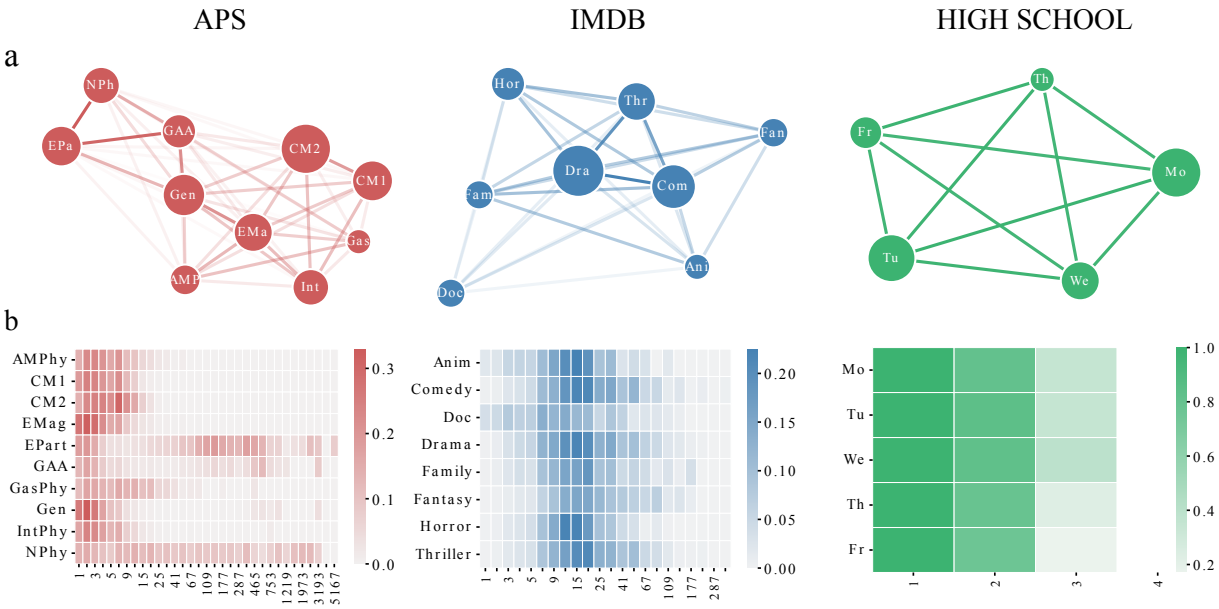

Figure 5.3: a) Each dataset is a graph in which vertices represent the layers of the multiplex hypergraphs and the thickness of an edge (α, β) quantifies the pairwise cosine similarity of layer activity matrices $\mathbf{B}_\alpha, \mathbf{B}_\beta$ associated with layers α and β. Vertex size is proportional to the number of nodes active in that layer. b) Matrix L associated with each dataset. Rows are normalized by the number of nodes active in each layer. Interaction orders are binned exponentially.

activity matrices, one for each layer α:

$$\mathbf{B}_\alpha = \{b_{id}\} = \begin{cases} 1 & \text{if } i \text{ is active at order } d \text{ in } \alpha \\ 0 & \text{otherwise.} \end{cases}$$

It can be useful to aggregate information about nodes and define an *aggregated layer activity matrix* L as:

$$L = \{l_{\alpha d}\} = |\{\text{nodes active at order } d \text{ in } \alpha\}|$$

In Fig. 5.3a, each dataset is represented as a graph where vertices are the layers of the multiplex hypergraph and links measure the similarity in activity patterns of two layers α and β, quantified as the cosine similarity of their node activity matrices $\mathbf{B}_\alpha$ and $\mathbf{B}_\beta$. The thicker the link, the higher the similarity. This figure emphasizes layers that not only share common active nodes, but also exhibit similar patterns of participation across different hyperedge orders. Particularly, a consistent higher-order similarity is observed across school days, reflecting recurring interaction patterns throughout the week. Other datasets show more heterogeneous behaviour, with documentary casts differing significantly from other layers, while drama and comedy casts exhibit similar patterns.

Fig. 5.3b shows the aggregated layer activity matrices L for the three datasets. To account for variations in layer size, we normalize each row by the total number of active nodes in the respective layer. Distinct collaboration patterns emerge across the subfields of physics and movie genres. For instance, scientists in General or Electromagnetic Physics usually contribute to papers with a smaller number of co-authors, whereas co-author groups in fields like Elementary Particles and Nuclear Physics exhibit more variation in size. In movie collaborations, actor activity is concentrated in medium-size groups, typically between ten and twenty members. However, documentaries often feature smaller casts, while family and comedy movies tend to have larger ones. The figure once again highlights how students at school maintain a consistent group size in their interactions throughout the week.

Similar to node activity, node degree (defined as the number of interactions in which a node participates) is another property that can be used to measure the activity across the different layers and interaction orders. We define a list $\mathbf{K}$ of *node degree matrices*, one for each node i:

$$\mathbf{K}_i = \{k_{\alpha d}\} = |\{\text{hyperedges of order } d \text{ involving } i \text{ at } \alpha\}|$$

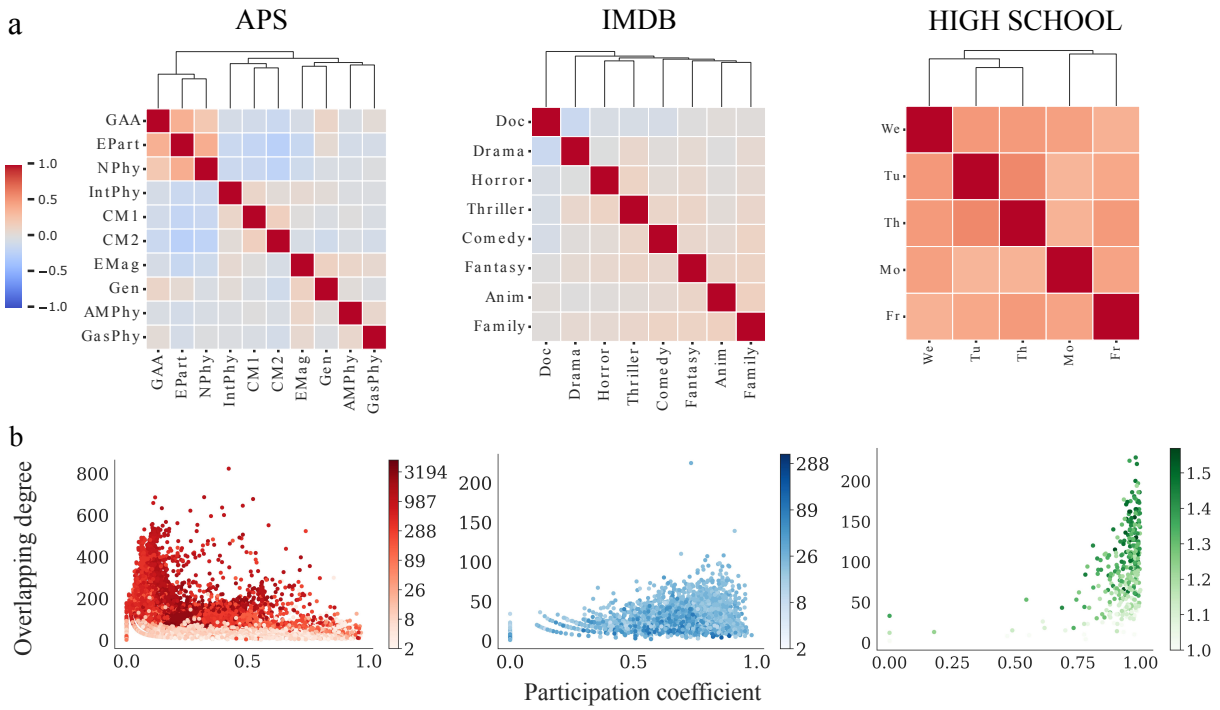

Figure 5.4: a) The heatmap shows the pairwise correlation between the degrees of nodes across different layers. The color scale indicates the strength of the correlation, with blue representing low correlation and red representing high correlation. b) A system unit i is represented as a point on a Cartesian plane, with the overlapping degree o_i on the y-axis, the participation coefficient P_i on the x-axis, and the average order of the interactions in which the unit is involved indicated by color intensity.

We use $k_{i\alpha}$ to denote the *total number of interactions* involving i in layer α, irrespective of their order:

$$k_{i\alpha} = \sum_{d=1}^{D} \mathbf{K}_{i\alpha d},$$

where D is the order of the largest interaction in the dataset.

In Fig. 5.4a, we analyze the correlation of node degree across layers, exploring the extent to which a node with a high or low degree in one layer similarly exhibits a high or low degree in another layer. The correlation matrix for physics collaborations uncovers a hierarchical structure, with strong correlations among specific subfields sharing commonalities and notable scientists, such as in Nuclear and Elementary Particles Physics. In contrast, the degree correlations among actors are generally weak, though certain genres, like thriller and horror, show similarities. A significant correlation in node degrees across consecutive days in HIGH SCHOOL suggests stable and structured daily interaction patterns, implying that individuals with numerous interactions on one day tend to maintain similar levels of interactions on subsequent days, and vice versa.

We now define the *overlapping degree* o_i for a node i as the total number of interactions involving i, irrespective of both layers or orders:

$$o_i = \sum_{\alpha=1}^{M} k_{i\alpha}$$

It can be interesting to measure (i) how the overlapping degree of a node i is spread across the layers, i.e., if the degree is concentrated in certain layers or if it is uniformly distributed; (ii) how interactions involving node i are spread across orders. We measure (i) by defining the *participation coefficient* P_i of a node i of the degree with respect to the layers:

$$P_i = \frac{M}{M-1}\left[1 - \sum_{\alpha=1}^{M}\left(\frac{k_{i\alpha}}{o_i}\right)^2\right]$$

where $k_{i\alpha}$ is the degree of node i at layer α, o_i is the overlapping degree of node i and M is the total number of layers. We measure (ii) by considering the average order of the interactions node i participates in.

In Fig. 5.4b, we represent each unit i of the different systems on a Cartesian plane, characterizing them across three distinct dimensions: their overlapping degree o_i (on the y-axis), their participation coefficient P_i (on the x-axis), and the average order of the interactions in which they are involved (indicated by color intensity). In general, such three dimensions provide different information about connectivity patterns and are only weakly

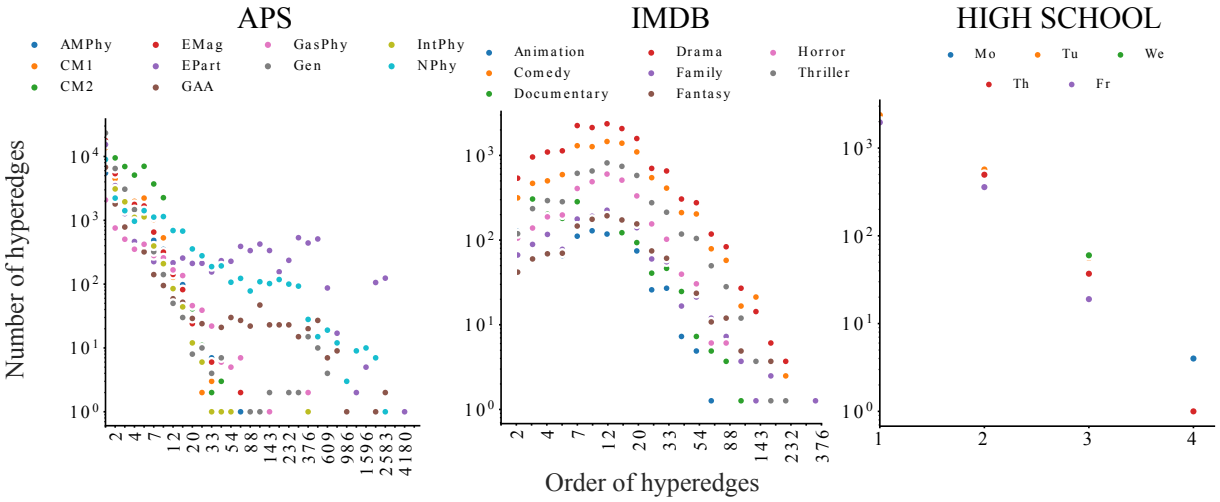

Figure 5.5: Distribution of hyperedge orders disaggregated by layers in each dataset. Colors distinguish between different layers, with interaction orders binned exponentially.

correlated, or even uncorrelated. In APS, scientists are spread across the plane in terms of degrees and average interaction order, displaying an overall tendency towards specialization in a selected number of physics subfields, yet the behavior remains heterogeneous. IMDB displays isolated outliers with a very high degree, low dispersion around the average interaction order, and a general tendency towards uniform participation across multiple genres. HIGH SCHOOL shows students covering the entire spectrum of node degrees and average group orders, with an expected tendency to interact uniformly across school days.

5.3 Hyperedge properties

We now turn our attention to the properties of the hyperedges encoding interactions in the same three real-world systems.

We begin by considering the simplest measure for characterizing higher-order interactions, namely, the order of the groups. In Fig. 5.5, we plot the hyperedge order distribution disaggregated by layers. APS and IMDB reveal heterogeneity across layers, suggesting that different physics subfields and movie genres exhibit distinct patterns of collaboration in terms of the number of people involved in a paper or a movie cast. For example, genres such as documentaries and animated movies typically feature fewer actors compared to other genres. Conversely, papers in Elementary Particles and Nuclear Physics often include a larger number of authors compared to those in other areas of physics. The distributions in HIGH SCHOOL are stable across layers, indicating that patterns of face-to-face interactions tend not to change over the days, with a general preference for smaller

groups over larger ones.

Another property frequently studied in the context of multiplex networks is edge overlap, which measures the extent to which interactions among the same nodes tend to repeat across multiple layers. We define *hyperedge overlap* as the maximum number of layers in which an interaction repeats exactly. In Fig. 5.6, we present the distribution of hyperedge overlap, including information about the order of the interactions. As expected, APS displays a high degree of hyperedge overlap, indicating that the same set of scientific authors consistently interact across multiple areas of physics. Conversely, for actors, hyperedge overlap decays very rapidly. Small interactions typically exhibit a higher degree of overlap than large interactions. Historically, edge overlap in multiplex networks with higher-order interactions has often been investigated by projecting hyperedges at different layers into cliques, frequently resulting in extremely high values of edge overlap. Our analysis suggests that patterns of hyperedge overlap are more complex and that projections of hyperedges can account for the high amount of overlap previously observed [22].

Finally, we assign a score P_e to each hyperedge e, defined in terms of the participation coefficient of the nodes involved in the interaction:

$$P_e = \frac{1}{|e|} \sum_{i \in e} P_i$$

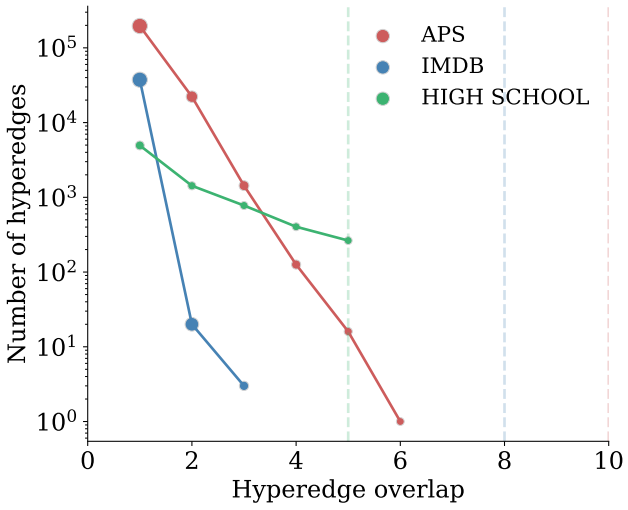

Figure 5.6: Number of hyperedges as a function of their overlap, i.e., the maximal number of layers in which an interaction repeats. Markers are scaled proportionally to the average order of hyperedges. Colored dashed lines indicate the corresponding number of layers in each dataset.

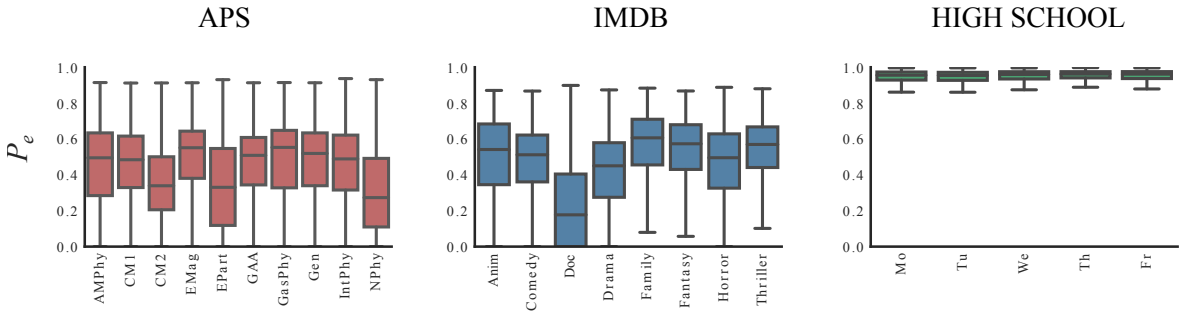

Figure 5.7: Boxplots showing the distribution of hyperedge participation coefficient P_e across layers in each dataset.

where $|e|$ represents the number of nodes participating in hyperedge e and P_i is the participation coefficient of node i, as defined in the previous section.

This measure captures the tendency of hyperedges to connect nodes that either specialize in a few layers or act as generalists across multiple layers. In Fig. 5.7, we use boxplots to show distributions of P for hyperedges in various layers. It is noteworthy that layers can display heterogeneous mean values for the participation coefficient of their hyperedges. For example, casts in documentaries and co-authors in Nuclear and Elementary Particles Physics tend to include specialists. On the other hand, family and thriller movies are more likely to feature generalist actors. In HIGH SCHOOL, layers exhibit a consistent maximum mean value for the participation coefficient of hyperedges, attributed to students' regular attendance at school each day.

5.4 Mesoscale properties

We finally shift our focus towards mesoscale structures, examining the emergence of communities and core-periphery structures within different layers of real-world hypergraphs.

Communities are groups of nodes that display a higher degree of connectivity among themselves than with the rest of the nodes in the system. In hypergraphs, a *community* is defined as a subset of nodes that tend to form cohesive units by participating in common hyperedges. When analyzing multiplex systems, it is typical to examine the similarities in community structures observed across various layers. In this direction, we employ a method for hard clustering, applied independently to each layer, which is an extension of the well-established Infomap algorithm to the case of hypergraphs [43]. In general, Infomap minimizes the map equation, which quantifies the description length required to represent the random walker's movements on the network [141]. This optimization effectively partitions the network into communities that best capture the inherent modular

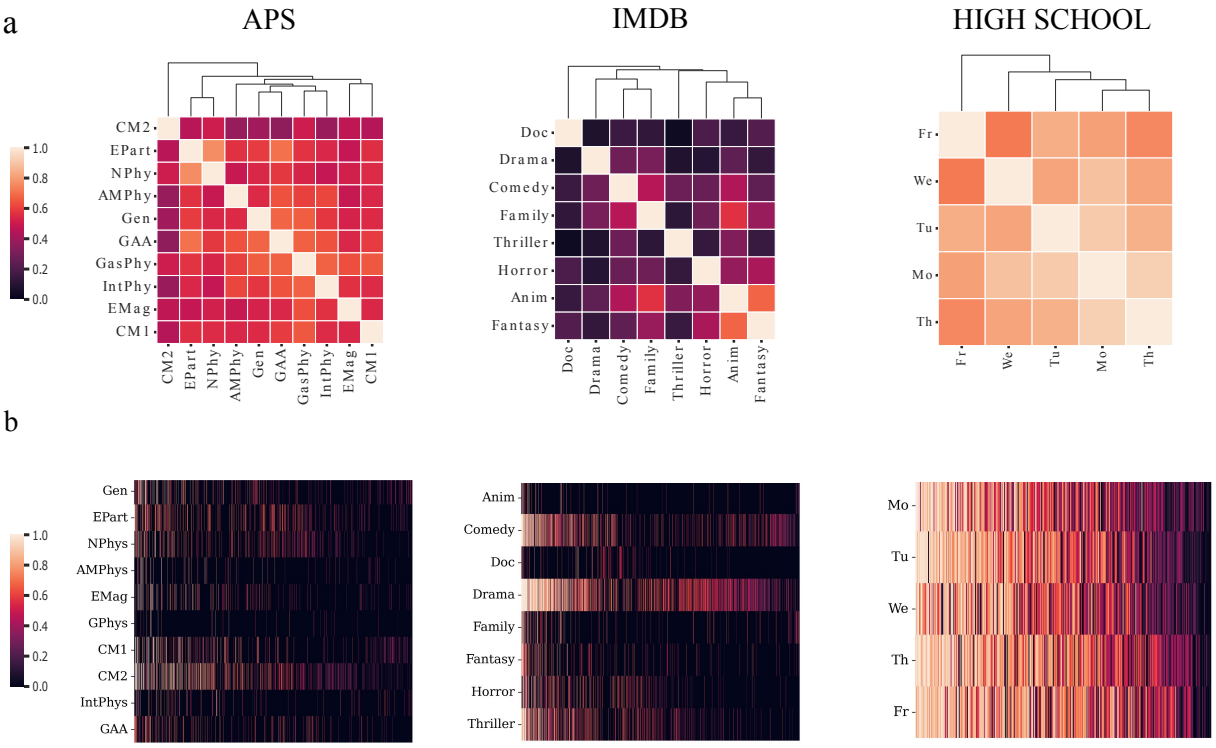

Figure 5.8: a) Heatmaps illustrating the similarity of community structures across layers, measured by Normalized Mutual Information (NMI), for the three datasets. In these heatmaps, high NMI values are represented by light colors, while low NMI values are represented by dark colors.

(b) Core-periphery score c_i across layers for each node in the datasets, visualized as heatmaps. Rows are layers and columns represent nodes. Lighter colors are higher coreness values. Nodes are consistently sorted across layers based on their coreness in the aggregated hypergraphs, i.e., the hypergraph obtained by collapsing all layers into a single layer.

structure. The method is publicly available [43] and we have used default parameters.

To assess the similarity and consistency of community structure across different hypergraph layers, we use Normalized Mutual Information (NMI), taking into account the set of nodes active in both layers. NMI is defined as:

$$\mathrm{NMI}(C_1, C_2) = \frac{I(C_1, C_2)}{\sqrt{H(C_1)H(C_2)}}$$

where $I(C_1, C_2)$ denotes the mutual information between partitions C_1 and C_2. Mutual information quantifies the amount of information shared between the two partitions, i.e., measures how much knowing the community structure in one partition informs about the structure in the other. $H(C_1)$ and $H(C_2)$ represent the entropies of the partitions C_1 and C_2, respectively. By normalizing the mutual information $I(C_1, C_2)$ with the geometric mean of these entropies, NMI adjusts for the variability in partition sizes and the number of communities, allowing for a fair comparison of community structures across different partitions. NMI ranges from 0 (indicating no mutual information) to 1 (indicating perfect agreement).

In Fig. 5.8a, we present the outcomes of this analysis through heatmaps, highlighting the strength of community structure similarities across layers. For example, community structures in animation and fantasy movie casts are closely related, as are those in comedy and family movie genres, whereas documentary casts show a completely uncorrelated structure. The community structure within HIGH SCHOOL interactions remains consistent across days. Physics collaborations reveal a significant degree of similarity across fields, though some layers exhibit more pronounced similarities than others. Overall, communities tend to be preserved within physics subfields and school days, while movie genres often demonstrate predominantly uncorrelated communities.

We then direct our attention to core-periphery structures. Core-periphery structures delineate the existence of a group of central and tightly connected nodes, the core, surrounded by less densely connected peripheral nodes, forming a distinctive organizational pattern often crucial for system functionality [142].

We detect core-periphery structure for each layer independently using a method tailored for hypergraphs [143]. Such a method ranks nodes assigning to each node i a value c_i, where c_i is a real number within the range of 0 to 1. This value delineates the extent to which a node participates in the core (value closer to 1) or the periphery (value closer to 0) structure of the system. Vector $\mathbf{c}$ is selected for each layer α independently. Following the work by Tudisco and Higham [143], for each layer α we select the vector $\mathbf{c}$ that optimizes

the following function:

$$\max_{\mathbf{c}} \sum_{e \in E_\alpha} \frac{1}{|e|} \cdot \left(\sum_{i \in e} c_i \right) \quad \text{subject to} \quad \|\mathbf{c}\|_2 = 1 \text{ and } c_i \geq 0 \text{ for all } i$$

This continuous scale allows for a nuanced characterization of each node's role within the core-periphery framework. Additionally, by comparing the coreness score c_i for each node i across the different layers of a multiplex hypergraph, we can analyze the variation or consistency of a node centrality across the layers. This method is publicly available [143]. We use the implementation from Hypergraphx [3].

To provide a visually appealing way of highlighting correlations of core-periphery structures and node behaviour across layers, Fig. 5.8b shows heatmaps in which rows are layers and columns are nodes, and each entry is coloured depending on node coreness. To compare the coreness value of single nodes across layers and visualize to which extent it keeps its core value, we maintain a consistent sorting of the nodes on the x-axis. For each dataset, nodes are sorted according to their core values in the aggregated hypergraphs (i.e., the hypergraph obtained by dropping information about layers and collapsing every hyperedge to a single layer). We observe that coreness values of nodes are maintained across layers exhibiting patterns similar to those seen in community structure correlations.

5.5 Discussion

Over time, novel and more comprehensive network models have emerged with the idea of capturing richer information about the interactions. In this Chapter we have introduced a general set of measures to characterize the structure of multiplex hypergraphs, which bridge the notion of (i) multiplex networks, describing links of different types, and (ii) hypergraphs, encoding non-dyadic ties. We introduced a description of nodes in terms of higher-order activity patterns and degrees, to quantify the extent and magnitude of node participation in interactions of different orders across layers. Nodes have been also characterized in terms of how their degree is correlated and spread in the different layers, and by their preferred order of group interactions. For hyperedges, we have studied their order distribution disaggregated by layers, highlighting different patterns of group interactions depending on the hyperedge type. We have quantified the extent to which hyperedge tends to repeat exactly in multiple layers and we have analyzed the layer-dependant property of hyperedges of linking nodes with low or high participation coefficients. Moreover, we have analyzed hypergraphs at their mesoscale, quantifying similarities of communities and

core-periphery participation across layers. Finally, we have validated our proposed measures on three datasets from different domains, describing collaboration patterns across physics subfields, movie genres and daily interactions among students and with the related randomized hypergraphs (see Appendix B.2), to assess the relevance of the results

In summary, we believe that these measures can be useful in describing the structure of various empirical datasets characterized by both multiplex and higher-order interactions. We also hope that this initial characterization of multiplex hypergraphs will spark interest from a methodological perspective, such as proposing frameworks for extracting multiplex communities in hypergraphs. Further characterizations could be enhanced by considering the complex patterns of temporality in hyperedges, a common feature of higher-order systems [8, 25, 144–148].

6 | Software and data for higher-order network analysis

In parallel with theoretical and methodological progress, a crucial role in advancing network science has been played by the development of efficient algorithms and computational tools to analyze networked data. Nowadays, widely used, community-based software such as NetworkX [149] and igraph [150], and individual efforts such as graph-tool [151] – 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 [152–156], in particular XGI [157], have started to develop computational tools to match the explosion of interest in higher-order systems. Another key factor driving the growth of network science is the availability of network datasets. These datasets are essential for developing and testing algorithms, validating theoretical models, exploring real-world patterns and phenomena, and enabling reproducibility. Efforts such as the Stanford Network Analysis Project (SNAP) [158], the Network Repository [159], the Koblenz Network Collection (KONECT) [90] and Netzschleuder [160] have played a major role in making network datasets accessible. However, most existing repositories collect data representing traditional networks, where edges can only connect two system units.

In this Chapter, we provide our contributions to expand the ecosystem of software and data for higher-order network analysis. In particular, we have developed Hypergraphx, a Python library for higher-order network analysis that offers accessible and efficient tools for constructing, visualizing, and analyzing hypergraphs. Complementing this project, we have created Hypergraph-data, a repository containing higher-order datasets with relational information and metadata from real-world systems across various domains. These contributions are aimed at lowering the entry barrier to higher-order network analysis and encouraging interdisciplinary collaboration.

6.1 Hypergraphx: a library for higher-order network analysis

Hypergraphx (HGX) is a multi-purpose, open-source, python library for the analysis of networked systems with 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. Hypergraphx is currently freely available on Github at `github.com/HGX-Team/hypergraphx`.

In this section, we discuss the main functionalities provided by HGX. The different tools of our library are illustrated online through detailed, user-friendly tutorials. Moreover, the library is conceived as an evolving, community-based effort, which will further extend its functionalities over the years.

Representations. Hypergraphs represent the most general and flexible framework to encode systems with higher-order interactions [23, 161]. 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 [23, 162] such as bipartite networks, maximal simplicial complexes, higher-order line graphs, dual hypergraphs and clique-expansion graphs.

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.

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 [129]. 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 subhypergraphs [32] and different centrality scores based on spectral approaches [33]. We also implement measures of hyperedge centrality based on shortest paths and betweenness flows [163].

Motifs. Motifs are small recurring patterns of subgraphs that are overrepresented in a network [60]. 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 [61]. We provide an implementation for higher-order motif analysis, extracting overabundant subgraphs of nodes connected by higher-order interactions, as originally defined in Ref. [1]. 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 [164].

Mesoscale structures. One of the most relevant features of graphs representing real-world systems is community structure [165]. 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 [166]. We also implement different generative models able to extract overlapping communities and jointly infer hyperedges [117], allowing to capture a variety of mesoscale organizations, including both disassortative and assortative community structure [167]. We provide a method able to extract hyperlink communities, where interactions, and not system units, are clustered across different hypergraph modules [4]. 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. [143].

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) [168]. 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 [45], and identifying significant maximally interacting sets, which represent the largest groups of nodes that interact significantly, captured by combining interactions of different orders [169].

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 [170, 171]. In our library, we offer ready-to-use implementations for various synthetic hypergraph samplers. We provide functions to build generalised Erdös-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 [39]. 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 [172]. Finally, we provide a higher-order activity-driven model with group interactions that change in time [173] and compute the associated percolation threshold.

Dynamical processes. The structural properties of complex networks shape the dynamical process occurring on top of them [174]. Recent works have revealed that higher-order interactions significantly impact various dynamical processes, including percolation [175], diffusion [46, 47], pattern formation [176, 177], synchronization [49–52, 178], contagion [53, 179, 180], and evolutionary games [56, 181, 182] 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 [51], to the implementation of the Master Stability Function approach for synchronization stability [52, 125]. We also provide an algorithm to simulate simplicial social contagion [53], and analytical and numerical tools to investigate random walks on hypergraphs [47].

Weighted, directed, signed, temporal and multiplex hypergraphs. Our library is highly flexible. It allows to store and analyze 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.

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.

6.2 Hypergraph-data: a repository for higher-order network data

Hypergraph-data is a repository of real-world hypergraph datasets, addressing the growing need for accessible, coherent, and heterogeneous higher-order data. This repository aims to serve as a useful resource to facilitate and support research in higher-order network analysis, complementing other existing repositories, including XGI-data [183], HypergraphRepository [184] and `www.cs.cornell.edu/~arb/data/`.

6.2.1 Data domains

The repository provides data on both natural and artificial systems with higher-order interactions across multiple domains. Many of these systems have already been central in previous studies focusing on the limitations of traditional pairwise models, advocating for higher-order network modeling as a more appropriate framework for their analysis.

In social systems, higher-order interactions are common in activities involving groups of individuals. The repository includes datasets on human face-to-face interactions collected from settings such as schools, hospitals, and conferences [14, 91, 101, 102, 119]. Co-authorship datasets describe collaboration efforts in research, where groups of authors synergistically contribute to the same publication [30, 93, 185]. Q&A forum datasets capture online interactions where users collectively engage with the same question, forming discussion groups. Voting data represent decision-making patterns involving groups of participants [86].

In technological systems, group interactions appear for example in communication and financial networks. In email datasets hyperedges model interactions encoding messages sent to multiple recipients [30, 186, 187]. Bitcoin transaction datasets record financial exchanges involving multiple senders and receivers in a single transaction, providing insights into decentralized financial systems.

Ecological systems also exhibit higher-order interactions, such as in animal proximity networks. These datasets track interactions among groups of animals in their natural habitats, offering valuable data for studying collective behavior and ecological dynamics [92].

In biology, higher-order interactions are often key to understanding complex systems. Gene-disease datasets connect groups of genes to specific conditions [100]. Drug association datasets capture groups of drugs that interact or share similar properties [30]. Metabolic reaction datasets represent groups of molecules involved in the same biochemical process, shedding light on metabolic pathways and their structure.

6.2.2 Data features

The repository hosts hypergraph datasets that encode both relational information, structural properties and metadata. In other words, each dataset provides not only the relationships among nodes encoded as a hypergraph structure, but also the contextual information associated with the hypergraph as a whole, as well as its edges and nodes.

The repository supports a wide range of hypergraph types. These include undirected hypergraphs, where simple hyperedges connect sets of nodes, suitable for contexts like group memberships or co-authorship networks. Directed hypergraphs extend the previous notion by specifying source and target nodes within hyperedges, making them ideal for modeling directional flows like information propagation or metabolic pathways. Weighted hypergraphs associated weights to hyperedges, quantifying the strength or importance of the relationships, useful in scenarios such as interaction frequencies in social networks. The repository also includes temporal hypergraphs, where hyperedges are associated with timestamps, capturing the evolution of relationships over time. These datasets are particularly relevant for studying dynamic systems, such as communication networks. Additionally, multiplex hypergraphs are supported, enabling multiple types of relationships between the same set of system units. In these cases, hyperedges are categorized into layers, each one representing a different type of interaction, such as distinct types of social ties.

In addition to structural features, the repository includes metadata, providing for each dataset a richer set of contextual information. At the system level, hypergraph metadata describes the dataset as a whole, including its name, domain, version, source, and global properties such as density or average hyperedge size. At the unit level, node metadata captures attributes of the entities the nodes represent, such as age or sex in a social network, or molecular types in a biological network. At the interaction level, hyperedges are annotated with attributes such as weights, timestamps, or domain-specific labels, such as the field of a scientific paper or the genre of a movie.

6.2.3 Data format

The repository uses two formats for storing datasets, balancing accessibility and computational efficiency. Hypergraphs are provided in a JSON format for ease of use, readability and compatibility, as well as in a specialized binarized format optimized for efficient processing with Hypergraphx [3].

JSON is a widely used and human-readable format particularly suited for representing networked data in a structured and self-descriptive manner. In the JSON format, nodes and hyperedges are stored as distinct objects. Nodes include attributes that describe the entities they represent, such as the name or sex of a high school student, while hyperedges list their component nodes and additional attributes such as weights, timestamps or categorical labels. Global metadata fields describe the dataset as a whole, including its name, version, domain and source.

While JSON excels in accessibility and interoperability, it is not optimized for the performance demands of large-scale datasets. To address this, the repository also provides a binarized format. By encoding hypergraph data as serialized Python arrays and dictionaries, this format minimizes storage overhead and significantly accelerates I/O operations. As a trade-off, we note that this format is compatible only with Hypergraphx, as it aligns closely with its internal data models.

6.2.4 Data sources and processing

A first major source of data is raw information extracted from publicly available websites and databases. For instance, we constructed hypergraphs sourcing from IMDB data, where nodes represent actors and hyperedges model co-participation in movies connecting them, and from ArXiv, where papers connect their co-authors as hyperedges. In converting raw data to hypergraphs, we preserved the original higher-order relationships without loss of information. In addition, we sourced several datasets from scientific publications. This collection ensures that the repository reflects the state-of-the-art in hypergraph research and includes benchmarks and examples that have been previously studied.

Another significant portion of the repository is data that has been recently re-discovered under the lens of higher-order networks. Several such datasets have originally been stored as bipartite graphs, which can be seamlessly converted into hypergraphs without loss of information. This is for instance the case of a variety of traditional data in ecology and biology. However, in some cases, data has actually been collected and stored as graphs, such as in the case of face-to-face interaction data for the SocioPattern project [188], where the presence of polyadic relationships is lost, as each group is decomposed into a fully connected clique of dyadic ties. For these datasets, reconstructing the original hypergraph structure requires additional information, such as fine-grained information about the temporality of each dyad, such that cliques formed by co-occurring temporal dyads can be encoded as hypergraphs, or specific inference.

For all the datasets, we retain information about the original source to ensure proper attribution and transparency. Each dataset includes metadata that credits the original work or source from which it was derived. Additionally, we provide clear citation guidelines for users, encouraging them to cite the original source.

6.2.5 Data versioning and integrity

The repository implements a system of semantic versioning and hash-based verification to effectively manage the evolution of datasets over time and ensure data reliability. This

approach addresses a critical challenge in the scientific community, since ensuring that shared datasets are accurately identified and remain consistent over time is fundamental for establishing transparent, fair and trustworthy benchmarks in research.

Our versioning system is built on the simple principles of semantic versioning, which assigns a version number to each dataset in the format MAJOR.MINOR.PATCH. This structured approach enables users to distinguish between significant changes, incremental updates, and minor corrections. We reserve major version changes to modifications that alter the structural properties of the networked systems, such as the addition or removal of new nodes or links. These changes may not maintain compatibility with previous versions, making clear identification essential. Minor updates reflect enhancements in the datasets, such as adding metadata to nodes and links, without altering the system structure. Patch updates address small corrections, such as fixing errors in node and link attributes. This versioning framework ensures transparency and allows users to consistently identify the exact dataset version and trace its evolution over time. It also guarantees that older versions remain accessible, supporting reproducibility in research. To complement the versioning system, detailed changelogs are maintained for each dataset, documenting the nature of changes and the specific aspects of the dataset that were updated.

In addition to semantic versioning, the repository employs a hash-based system to verify dataset versions and ensure data integrity. Each version of a dataset is associated with a unique cryptographic hash (SHA-256). The hash serves as a digital fingerprint of the dataset. This approach allows users to independently verify the dataset they are working with by comparing the hash of the downloaded file to the recorded hash in the repository. A matching hash confirms that the dataset is unaltered and corresponds to the intended version. Any modification to the dataset results in a mismatched hash, signalling potential integrity issues.

6.2.6 Repository website

The repository website is designed to be an intuitive interface for users to explore and download hypergraph datasets. The homepage provides an overview of the purpose and scope of the repository, with quick access to data sets and updates, such as recently added data or version changes. Users can navigate the dataset browser to explore the full data catalogue. The catalogue includes search and filtering tools, allowing users to narrow their options based on specific criteria. Datasets can be filtered by domain, such as social, biological, or technological networks, or by dataset characteristics, such as the number of nodes and the number of hyperedges, or by structural features including the presence of

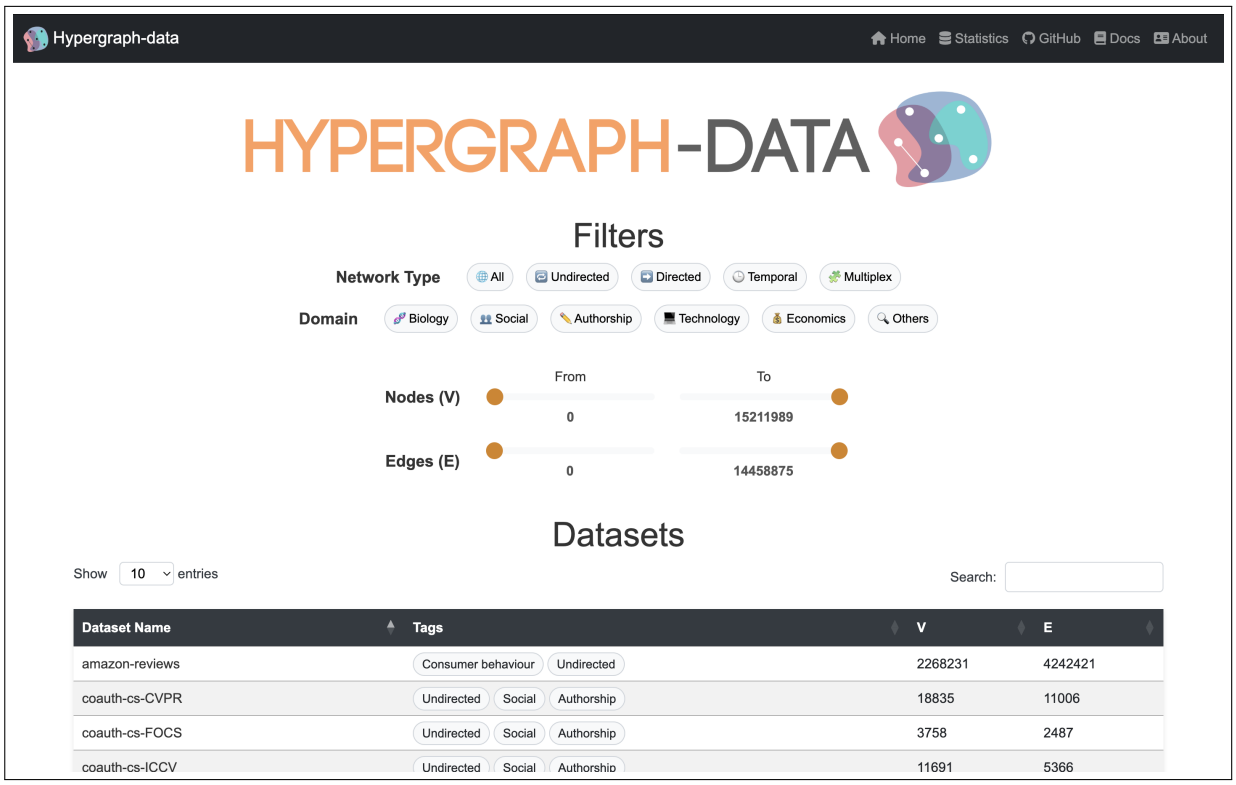

Figure 6.1: Home page of the repository website, showcasing its interface and key features for navigating and accessing datasets.

temporal annotations or weighted hyperedges. In Fig. 6.1, we present the homepage of the website associated with our data repository.

Each dataset is associated with a dedicated page that provides more details about the system, including a comprehensive overview of its metadata and characteristics. On these pages, users can preview key attributes of the selected dataset, such as its structural summary statistics, before downloading it. Moreover, such pages contain also information about current and older versions and supported formats, as well as appropriate source references in BibTex format. To offer a quick and informative snapshot of the dataset, we also provide basic data visualizations, such as hyperedge size distributions or temporal activity graphs, directly on the detail pages.

6.3 A guided tour of HGX using real-world data

In order to illustrate the power of HGX in loading, manipulating, analysing and visualizing real-world systems with group interactions, in Figure 6.2 we present an illustrative analysis of a dataset from the SocioPattern collaboration encoding face-to-face social interactions in a high school [14]. This dataset has been widely investigated in the literature on higher-order interactions [1, 30, 53, 117, 172]. We recall that it 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 (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 (B) we perform higher-order motif analysis as introduced in Ref. [1]. 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 [39]. Local structures with group interactions supported by pairwise links are found to be particularly relevant. In (C) we describe the mesoscale structure of the system, by extracting overlapping communities with the method of Ref. [117]. 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 (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 [45], removing redundant hyperedges and identifying the most informative group interactions. We continue by showcasing the ability of the model introduced in Ref. [172] to generate hypergraphs which are similar to the original dataset. To validate such a statement, in (E) we plot the distribution of (a rescaled version of) higher-order centrality measure [32] 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 (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. [145]. 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 [53]. In (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 (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 [45] among two, three and four of them.

6.4 Discussion

Advances in data-driven research fields are closely tied to the development of robust software tools and the maintenance of open and reliable data repositories.

Here we have presented HGX, a versatile and robust python library that offers a flexible and efficient framework to analyze 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. Moreover, we aim to expand our coverage of higher-order processes, by including different evolutionary games [56, 182], ecological dynamics [29], and more.

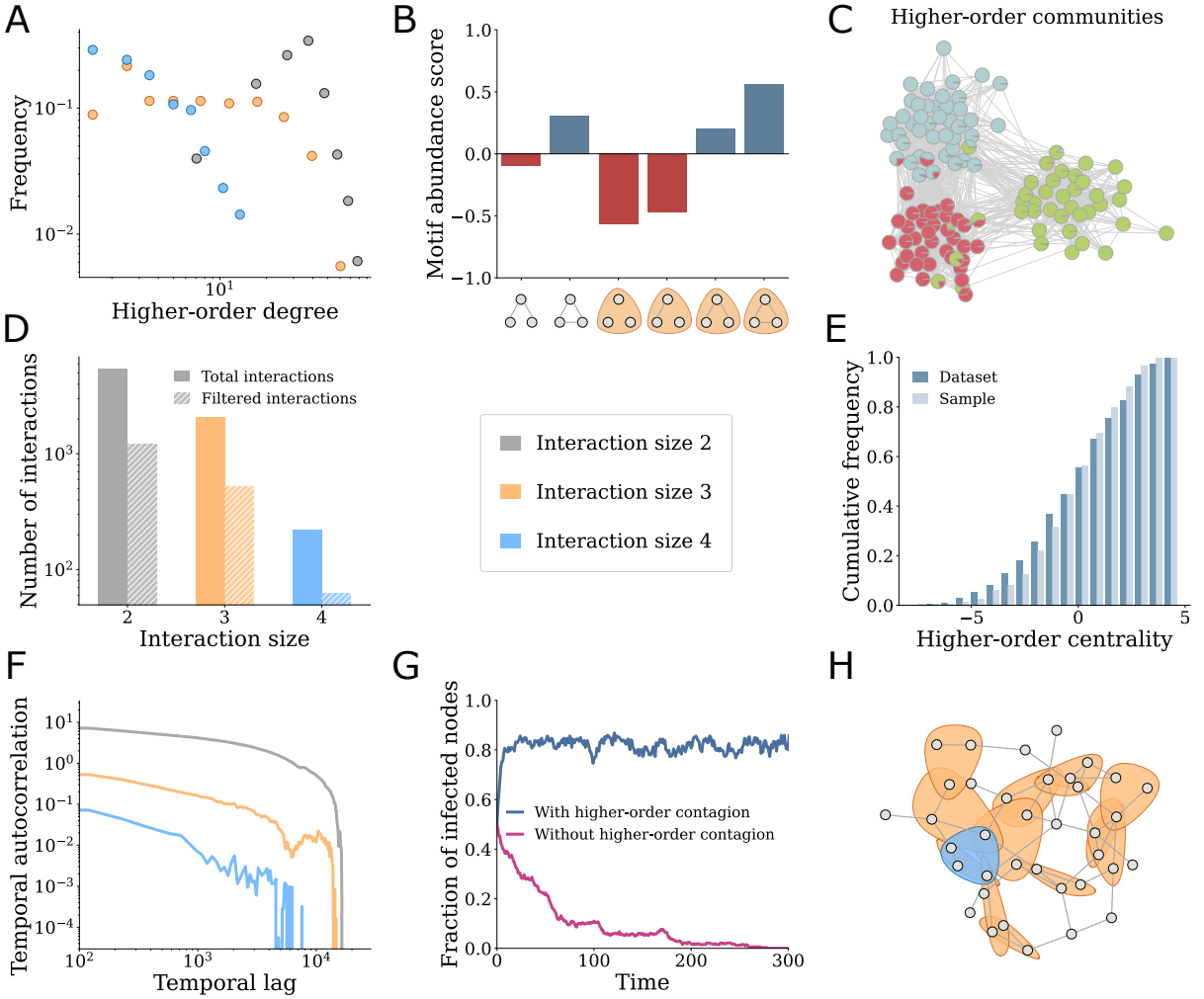

Figure 6.2: **Higher-order analysis of social interactions.** We illustrate different functionalities of HGX on a dataset of face-to-face group interactions in a school from the SocioPattern collaboration [14]. (A) Higher-order degree distributions for different interaction sizes. (B) Higher-order motif 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).

We also have introduced Hypergraph-data, a centralized repository specifically designed for hypergraphs, addressing the growing need for an organized and diverse collection of datasets capturing higher-order relationships. Complementing existing sources focusing on pairwise network data, our repository fills a critical gap in the network science community, giving higher-order network scientists access to a large, open and standardized amount of empirical data, enabling the testing of new ideas, and their reproducibility.

Future expansions of Hypergraph-data could include tools for interactive visualization of hypergraphs and their associated statistics, enabling researchers to explore and analyze higher-order data directly on the website. Looking ahead, we also foresee community efforts to establish a unified and well-documented JSON standard for hypergraph data formats. Such a standard would enable seamless integration across libraries and programming languages, reducing barriers for researchers and fostering more efficient collaboration.

In summary, HGX and hypergraph-data aim to provide the research community with essential resources to explore and analyze higher-order systems, advancing the development of network science beyond traditional dyadic approaches.

7 | Conclusions

Networks have emerged as a powerful framework for modeling and analyzing complex systems across multiple domains, from sociology to biology. Despite the simplicity of early network models, which encoded relations as simple, unweighted and undirected binary edges, researchers were able to extract deep insights into the organization of real-world systems, including small-world properties, scale-free degree distributions and clustering phenomena. Over time, traditional network representations proved too limited to fully describe real-world interactions. Weighted networks introduced varying connection strengths, while directed networks incorporated asymmetries of information flow. Temporal networks captured the evolution of relationships and multiplex networks enabled the representation of different types of relationships among the same entities. While these developments reflected a step toward capturing the rich complexity found in real-world systems, traditional frameworks remain fundamentally limited by their focus on pairwise interactions. Real-world systems, from ecological food webs to social groups, are dominated by many-body interactions, that cannot be properly described by pairwise interactions. To address this limitation, researchers have turned to higher-order network models, like hypergraphs and simplicial complexes, to encode group interactions accurately. These models represent the next logical step in the evolution of network science, offering the potential to explore complex systems more comprehensively and extract richer insights into their structure and dynamics.

In this Thesis, we have proposed several methodological and computational tools to investigate structural properties of hypergraphs, advancing our ability to extract meaningful insights from higher-order data. First, we have developed tools for motif analysis, introducing algorithms to identify statistically over-represented patterns of higher-order interactions. By comparing motif frequencies to suitable null models, we have described higher-order networks at their microscale, revealing the emergence of distinct families of hypergraphs characterized by similar local connectivity patterns. To overcome the computational limitations of exact motif analysis in large systems, we have proposed a sampling method which improves scalability while maintaining high accuracy. We have then investigated hypergraphs at their mesoscale through hyperlink communities, introducing a

framework that naturally captures both the hierarchical organization of higher-order interactions and community overlap. We turned our attention to generalized hypergraphs, to accommodate and study group interactions described by more complex features. For directed hypergraphs, we have introduced measures of reciprocity and methods to analyze directed group interactions, providing insights into systems like metabolic networks and financial transactions. We have also developed tools for multiplex hypergraphs, enabling the analysis of systems where nodes participate in different types of group interactions simultaneously, including scientific collaborations across different fields. To tackle the crucial bottleneck in higher-order network science of the lack of accessible, scalable, and well-documented computational tools and standardized datasets, we have developed Hypergraphx, an open-source Python library for higher-order networks. This software provides accessible and efficient tools for hypergraph construction, visualization and analysis. We have complemented this project with hypergraph-data, a repository of datasets with relational information and metadata of real-world systems spanning different domains. These contributions aim to lower the entry barrier for higher-order network analysis and foster collaborations across disciplines.

Looking ahead, several promising research directions could extend the methods developed in this thesis. The increasing availability of large-scale higher-order data calls for more sophisticated algorithmic solutions. While our sampling methods proved effective for motif analysis, similar approaches could be developed for other computationally intensive tasks in hypergraph analysis, such as community detection or other pattern mining tasks. The theoretical framework for hypergraphs could be expanded with new tools and measures to better characterize systems with higher-order interactions, possibly with direction or with temporal and multiplex features. On the software side, Hypergraphx could be extended to support massive hypergraphs with distributed computing and GPU acceleration. The development of visualization tools specifically designed for higher-order networks would also facilitate data exploration and result interpretation. Additionally, establishing benchmark datasets and standardized evaluation metrics would help assess and compare different methods for hypergraphs.

As higher-order network science continues to mature, it offers the chance to unlock new insights into phenomena that have been missed or oversimplified by traditional binary approaches. By fostering interdisciplinary collaboration and bridging the gap between theory and applications, we can move closer to realizing the full potential of higher-order networks, hopefully transforming our ability to model, analyze and ultimately understand the complexity of the world around us.

Bibliography

[1] Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "Higher-order motif analysis in hypergraphs," *Communications Physics*, vol. 5, no. 1, p. 79, 2022.

[2] Q. F. Lotito, F. Musciotto, F. Battiston, and A. Montresor, "Exact and sampling methods for mining higher-order motifs in large hypergraphs," *Computing*, vol. 106, no. 2, pp. 475–494, 2024.

[3] 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*, vol. 11, no. 3, p. cnad019, 2023.

[4] Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "Hyperlink communities in higher-order networks," *Journal of Complex Networks*, vol. 12, no. 2, p. cnae013, 2024.

[5] Q. F. Lotito, A. Montresor, and F. Battiston, "Multiplex measures for higher-order networks," *Applied Network Science*, vol. 9, no. 1, p. 55, 2024.

[6] Q. F. Lotito, A. Vendramini, A. Montresor, and F. Battiston, "The microscale organization of directed hypergraphs," *arXiv preprint arXiv:2410.16258*, 2024.

[7] Q. F. Lotito, L. Betti, B. Nortier, A. Montresor, and F. Battiston, "Hypergraph-data: a repository for higher-order network data," *Working paper*, 2025.

[8] B. Arregui-García, A. Longa, Q. F. Lotito, S. Meloni, and G. Cencetti, "Patterns in temporal networks with higher-order egocentric structures," *Entropy*, vol. 26, no. 3, p. 256, 2024.

[9] S. Genetti, E. Ribaga, E. Cunegatti, Q. F. Lotito, and G. Iacca, "Influence maximization in hypergraphs using multi-objective evolutionary algorithms," in *International Conference on Parallel Problem Solving from Nature*, pp. 217–235, Springer, 2024.

[10] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex net-

works: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[11] M. E. J. Newman, *Networks: An Introduction*. Oxford; New York: Oxford University Press, 2010.

[12] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75–174, 2010.

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[14] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PloS one*, vol. 10, no. 9, p. e0136497, 2015.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[16] P. Erdős, "On random graphs I," *Publ. Math.*, vol. 6, pp. 290–297, 1959.

[17] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.

[19] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Physical review letters*, vol. 93, no. 26, p. 268701, 2004.

[20] P. Holme and J. Saramäki, "Temporal networks," *Phys. Rep.*, vol. 519, no. 3, pp. 97–125, 2012.

[21] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, "Mathematical formulation of multilayer networks," *Phys. Rev. X*, vol. 3, p. 041022, Dec 2013.

[22] F. Battiston, V. Nicosia, and V. Latora, "Structural measures for multiplex networks," *Physical Review E*, vol. 89, no. 3, p. 032804, 2014.

[23] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, "Networks beyond pairwise interactions: structure and dynamics," *Physics Reports*, vol. 874, pp. 1–92, 2020.

[24] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, *et al.*, "The physics of higher-order interactions in complex systems," *Nature Physics*, vol. 17, no. 10, pp. 1093–1098, 2021.

[25] G. Cencetti, F. Battiston, B. Lepri, and M. Karsai, "Temporal properties of higher-order interactions in social networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[26] A. Patania, G. Petri, and F. Vaccarino, "The shape of collaborations," *EPJ Data Science*, vol. 6, pp. 1–16, 2017.

[27] S. Klamt, U.-U. Haus, and F. Theis, "Hypergraphs and cellular networks," *PLOS Computational Biology*, vol. 5, no. 5, p. e1000385, 2009.

[28] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino, "Homological scaffolds of brain functional networks," *Journal of The Royal Society Interface*, vol. 11, no. 101, p. 20140873, 2014.

[29] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," *Nature*, vol. 548, no. 7666, pp. 210–213, 2017.

[30] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proceedings of the National Academy of Sciences*, vol. 115, no. 48, pp. E11221–E11230, 2018.

[31] S. Ranshous, C. A. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters, "Exchange pattern mining in the bitcoin transaction directed hypergraph," in *Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21*, pp. 248–263, Springer, 2017.

[32] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality and clustering in complex hyper-networks," *Phys. A*, vol. 364, pp. 581–594, 2006.

[33] A. R. Benson, "Three hypergraph eigenvector centralities," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 293–312, 2019.

[34] P. Chodrow and A. Mellor, "Annotated hypergraphs: models and applications," *Applied network science*, vol. 5, no. 1, p. 9, 2020.

[35] H. Yin, A. R. Benson, and J. Leskovec, "Higher-order clustering in networks," *Phys. Rev. E*, vol. 97, p. 052306, May 2018.

[36] N. Veldt, A. R. Benson, and J. Kleinberg, "Higher-order homophily is combinatorially impossible," *arXiv preprint arXiv:2103.11818*, 2021.

[37] J.-G. Young, G. Petri, and T. P. Peixoto, "Hypergraph reconstruction from network data," *arXiv preprint arXiv:2008.04948*, 2020.

[38] O. T. Courtney and G. Bianconi, "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes," *Physical Review E*, vol. 93, no. 6, p. 062311, 2016.

[39] P. S. Chodrow, "Configuration models of random hypergraphs," *Journal of Complex Networks*, vol. 8, no. 3, p. cnaa018, 2020.

[40] K. Kovalenko, I. Sendiña-Nadal, N. Khalil, A. Dainiak, D. Musatov, A. M. Raigorodskii, K. Alfaro-Bittner, B. Barzel, and S. Boccaletti, "Growing scale-free simplices," *Communications Physics*, vol. 4, no. 1, pp. 1–9, 2021.

[41] A. P. Millán, R. Ghorbanchian, N. Defenu, F. Battiston, and G. Bianconi, "Local topological moves determine global diffusion properties of hyperbolic higher-order networks," *Physical Review E*, vol. 104, no. 5, p. 054302, 2021.

[42] T. Carletti, D. Fanelli, and R. Lambiotte, "Random walks and community detection in hypergraphs," *Journal of Physics: Complexity*, vol. 2, no. 1, p. 015011, 2021.

[43] A. Eriksson, D. Edler, A. Rojas, M. de Domenico, and M. Rosvall, "How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs," *Communications Physics*, vol. 4, no. 1, pp. 1–12, 2021.

[44] P. S. Chodrow, N. Veldt, and A. R. Benson, "Generative hypergraph clustering: From blockmodels to modularity," *Science Advances*, vol. 7, no. 28, p. eabh1303, 2021.

[45] F. Musciotto, F. Battiston, and R. N. Mantegna, "Detecting informative higher-order interactions in statistically validated hypergraphs," *Communications Physics*, vol. 4, no. 1, pp. 1–9, 2021.

[46] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie, "Random walks on simplicial complexes and the normalized hodge 1-laplacian," *SIAM Review*, vol. 62, no. 2, pp. 353–391, 2020.

[47] T. Carletti, F. Battiston, G. Cencetti, and D. Fanelli, "Random walks on hypergraphs," *Physical Review E*, vol. 101, no. 2, p. 022308, 2020.

[48] C. Bick, P. Ashwin, and A. Rodrigues, "Chaos in generically coupled phase oscillator networks with nonpairwise interactions," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 26, no. 9, p. 094814, 2016.

[49] P. S. Skardal and A. Arenas, "Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching," *Communications Physics*, vol. 3, no. 1, pp. 1–6, 2020.

[50] A. P. Millán, J. J. Torres, and G. Bianconi, "Explosive higher-order kuramoto dynamics on simplicial complexes," *Physical Review Letters*, vol. 124, no. 21, p. 218301, 2020.

[51] M. Lucas, G. Cencetti, and F. Battiston, "Multiorder laplacian for synchronization in higher-order networks," *Physical Review Research*, vol. 2, no. 3, p. 033410, 2020.

[52] L. V. Gambuzza, F. Di Patti, L. Gallo, S. Lepri, M. Romance, R. Criado, M. Frasca, V. Latora, and S. Boccaletti, "Stability of synchronization in simplicial complexes," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.

[53] I. Iacopini, G. Petri, A. Barrat, and V. Latora, "Simplicial models of social contagion," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[54] S. Chowdhary, A. Kumar, G. Cencetti, I. Iacopini, and F. Battiston, "Simplicial contagion in temporal higher-order networks," *Journal of Physics: Complexity*, vol. 2, no. 3, p. 035019, 2021.

[55] L. Neuhäuser, M. T. Schaub, A. Mellor, and R. Lambiotte, "Opinion dynamics with multi-body interactions," *arXiv preprint arXiv:2004.00901*, 2020.

[56] U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, and V. Latora, "Evolutionary dynamics of higher-order interactions in social networks," *Nature Human Behaviour*, vol. 5, no. 5, pp. 586–595, 2021.

[57] G. Ghoshal, V. Zlatić, and G. Caldarelli, "Random hypergraphs and their applications," *Physical Review E*, vol. 79, no. 6, p. 066118, 2009.

[58] S. Citraro, J. Warner-Willich, F. Battiston, C. S. Siew, G. Rossetti, and M. Stella, "Hypergraph models of the mental lexicon capture greater information than pairwise networks for predicting language learning," *New Ideas in Psychology*, vol. 71, p. 101034, 2023.

[59] M. Lucas, L. Gallo, A. Ghavasieh, F. Battiston, and M. De Domenico, "Functional reducibility of higher-order networks," *arXiv preprint arXiv:2404.08547*, 2024.

[60] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[61] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.

[62] U. Alon, "Network motifs: Theory and experimental approaches," *Nat. Rev. Genet.*, vol. 8, no. 6, p. 450, 2007.

[63] K. Juszczyszyn, P. Kazienko, and K. Musiał, "Local topology of social network based on motif analysis," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 97–105, Springer, 2008.

[64] X. Hong-lin, Y. Han-bing, G. Cui-fang, and Z. Ping, "Social network analysis based on network motifs," *Journal of Applied Mathematics*, vol. 2014, 2014.

[65] O. Sporns and R. Kötter, "Motifs in brain networks," *PLoS Biol*, vol. 2, no. 11, p. e369, 2004.

[66] A. K. Dey, Y. R. Gel, and H. V. Poor, "What network motifs tell us about resilience and reliability of complex networks," *Proceedings of the National Academy of Sciences*, vol. 116, no. 39, pp. 19368–19373, 2019.

[67] A. C. Schwarze and M. A. Porter, "Motifs for processes on networks," *ArXiv*, vol. abs/2007.07447, 2020.

[68] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nat. Genet.*, vol. 31, no. 1, p. 64, 2002.

[69] A. Mazurie, S. Bottani, and M. Vergassola, "An evolutionary and functional assessment of regulatory network motifs," *Genome Biology*, vol. 6, no. 4, pp. 1–12, 2005.

[70] R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai, "Aggregation of topological motifs in the escherichia coli transcriptional regulatory network," *BMC bioinformatics*, vol. 5, no. 1, pp. 1–8, 2004.

[71] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon,

and H. Margalit, "Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction," *Proceedings of the National Academy of Sciences*, vol. 101, no. 16, pp. 5934–5939, 2004.

[72] L. Chen, X. Qu, M. Cao, Y. Zhou, W. Li, B. Liang, W. Li, W. He, C. Feng, X. Jia, *et al.*, "Identification of breast cancer patients based on human signaling network motifs," *Scientific reports*, vol. 3, no. 1, pp. 1–7, 2013.

[73] S. Patra and A. Mohapatra, "Review of tools and algorithms for network motif discovery in biological networks," *IET Systems Biology*, vol. 14, no. 4, pp. 171–189, 2020.

[74] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini, "Detecting early signs of the 2007–2008 crisis in the world trade," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[75] J. Bascompte and D. B. Stouffer, "The assembly and disassembly of ecological networks," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1524, pp. 1781–1787, 2009.

[76] B. I. Simmons, A. R. Cirtwill, N. J. Baker, H. S. Wauchope, L. V. Dicks, D. B. Stouffer, and W. J. Sutherland, "Motifs in bipartite ecological networks: uncovering indirect interactions," *Oikos*, vol. 128, no. 2, pp. 154–170, 2019.

[77] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.

[78] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[79] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Phys. Rev. E*, vol. 71, p. 065103, Jun 2005.

[80] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, "Temporal motifs in time-dependent networks," *J. Stat. Mech. Theory Exp.*, vol. 2011, no. 11, p. P11005, 2011.

[81] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 601–610, ACM, 2017.

[82] M. Kivelä and M. A. Porter, "Isomorphisms in multilayer networks," *IEEE Transactions on Network Science and Engineering*, vol. 5, no. 3, pp. 198–211, 2018.

[83] F. Battiston, V. Nicosia, M. Chavez, and V. Latora, "Multilayer motif analysis of brain networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 4, p. 047404, 2017.

[84] G. Lee, J. Ko, and K. Shin, "Hypergraph motifs: concepts, algorithms, and discoveries," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2256–2269, 2020.

[85] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 3, p. 347–359, Oct. 2006.

[86] L. Epstein, T. G. Walker, N. S. S. Hendrickson, and J. Roberts, "The U.S. Supreme Court Justices Database," 2019.

[87] `https://journals.aps.org/datasets`, 2021.

[88] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PLOS ONE*, vol. 10, no. 9, pp. 1–26, 2015.

[89] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Trans. on Mobile Comput.*, vol. 6, no. 6, pp. 606–620, 2007.

[90] J. Kunegis, "KONECT – The Koblenz Network Collection," in *Proc. Int. Conf. on World Wide Web Companion*, pp. 1343–1350, 2013.

[91] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?," *EPJ Data Science*, vol. 7, p. 11, May 2018.

[92] V. Gelardi, J. Godard, D. Paleressompoulle, N. Claidiere, and A. Barrat, "Measuring social networks in primates: Wearable sensors versus direct observations," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, p. 20190737, 04 2020.

[93] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of microsoft academic service (MAS) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ACM Press, 2015.

[94] J. Leskovec, D. Huttenlocher, and J. Kleinberg, *Signed Networks in Social Media*, p. 1361–1370. New York, NY, USA: Association for Computing Machinery, 2010.

[95] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), p. 641–650, Association for Computing Machinery, 2010.

[96] J. Piñero, J. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, 11 2019.

[97] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, pp. D833–D839, 10 2016.

[98] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: A discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 01 2015.

[99] N. Queralt-Rosinach, J. Piñero, A. Serrano, F. Sanz, and L. I. Furlong, "Disgenet-rdf: harnessing the innovative power of the semantic web to explore the genetic basis of diseases," 11 2015.

[100] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, "Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases," *PLOS ONE*, vol. 6, pp. 1–13, 06 2011.

[101] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, "Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers," *Network Science*, vol. 3, no. 3, pp. 326–347, 2015.

[102] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, and N. Voirin, "Estimating potential infection transmission routes in hospital wards using wearable proximity sensors," *PloS one*, vol. 8, no. 9, p. e73970, 2013.

[103] Q. F. Lotito, "Higher-order motif discovery sampling algorithm," 2022.

[104] Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "Higher-order motif analysis in hypergraphs," *Communications Physics*, vol. 5, no. 1, p. 79, 2022.

[105] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics reports*, vol. 659, pp. 1–44, 2016.

[106] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[107] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[108] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, "Extracting the hierarchical organization of complex systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 39, pp. 15224–15229, 2007.

[109] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[110] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[111] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007.

[112] T. S. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," *Phys. Rev. E*, vol. 80, p. 016105, Jul 2009.

[113] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[114] M. M. Wolf, A. M. Klinvex, and D. M. Dunlavy, "Advantages to modeling relational data using hypergraphs versus graphs," in *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–7, IEEE, 2016.

[115] A. Vazquez, "Finding hypergraph communities: a bayesian approach and variational solution," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 07, p. P07006, 2009.

[116] P. Chodrow, N. Eikmeier, and J. Haddock, "Nonbacktracking spectral clustering of nonuniform hypergraphs," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 2, pp. 251–279, 2023.

[117] M. Contisciani, F. Battiston, and C. De Bacco, "Inference of hyperedges and overlapping communities in hypergraphs," *Nature communications*, vol. 13, no. 1, pp. 1–10, 2022.

[118] N. Ruggeri, F. Battiston, and C. De Bacco, "Framework to generate hypergraphs with community structure," *Physical Review E*, vol. 109, no. 3, p. 034309, 2024.

[119] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, *et al.*, "High-resolution measurements of face-to-face contact patterns in a primary school," *PloS one*, vol. 6, no. 8, p. e23176, 2011.

[120] Q. F. Lotito, F. Musciotto, A. Montresor, and F. Battiston, "hyperlink-communities." `https://github.com/FraLotito/hyperlink-communities`.

[121] R. Guimerà and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, 2005.

[122] P. Traversa, G. Ferraz de Arruda, A. Vazquez, and Y. Moreno, "Robustness and complexity of directed and weighted metabolic hypergraphs," *Entropy*, vol. 25, no. 11, p. 1537, 2023.

[123] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, "Directed hypergraphs and applications," *Discrete Applied Mathematics*, vol. 42, no. 2, pp. 177–201, 1993.

[124] G. Preti, A. Fazzone, G. Petri, and G. De Francisci Morales, "Higher-order null models as a lens for social systems," *Physical Review X*, vol. 14, no. 3, p. 031032, 2024.

[125] L. Gallo, R. Muolo, L. V. Gambuzza, V. Latora, M. Frasca, and T. Carletti, "Synchronization induced by directed higher-order interactions," *Communications Physics*, vol. 5, no. 1, p. 263, 2022.

[126] H. Moon, H. Kim, S. Kim, and K. Shin, "Four-set hypergraphlets for characterization of directed hypergraphs," *arXiv preprint arXiv:2311.14289*, 2023.

[127] N. Pearcy, J. J. Crofts, and N. Chuzhanova, "Hypergraph models of metabolism," *International Journal of Biological, Veterinary, Agricultural and Food Engineering*, vol. 8, no. 8, pp. 752–756, 2014.

[128] S. Kim, M. Choe, J. Yoo, and K. Shin, "Reciprocity in directed hypergraphs: measures, findings, and generators," *Data Mining and Knowledge Discovery*, vol. 37, no. 6, pp. 2330–2388, 2023.

[129] S. Wasserman, K. Faust, *et al.*, *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[130] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E*, vol. 66, p. 035101, Sep 2002.

[131] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, pp. 626–688, 2015.

[132] S. Feng, E. Heath, B. Jefferson, C. Joslyn, H. Kvinge, H. D. Mitchell, B. Praggastis, A. J. Eisfeld, A. C. Sims, L. B. Thackray, *et al.*, "Hypergraph models of biological networks to identify genes critical to pathogenic viral response," *BMC bioinformatics*, vol. 22, no. 1, p. 287, 2021.

[133] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics reports*, vol. 544, no. 1, pp. 1–122, 2014.

[134] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[135] M. Barigozzi, G. Fagiolo, and D. Garlaschelli, "Multinetwork of international trade: A commodity-specific analysis," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 81, no. 4, p. 046104, 2010.

[136] A. Aleta, S. Meloni, and Y. Moreno, "A multilayer perspective for the analysis of urban transportation systems," *Scientific reports*, vol. 7, no. 1, p. 44359, 2017.

[137] F. Battiston, J. Iacovacci, V. Nicosia, G. Bianconi, and V. Latora, "Emergence of multiplex communities in collaboration networks," *PloS one*, vol. 11, no. 1, p. e0147451, 2016.

[138] M. De Domenico, "Multilayer modeling and analysis of human brain networks," *Giga Science*, vol. 6, no. 5, p. gix004, 2017.

[139] H. Sun and G. Bianconi, "Higher-order percolation processes on multiplex hypergraphs," *Physical Review E*, vol. 104, no. 3, p. 034306, 2021.

[140] V. Nicosia and V. Latora, "Measuring and modeling correlations in multiplex networks," *Phys. Rev. E*, vol. 92, p. 032805, Sep 2015.

[141] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.

[142] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, no. 4, pp. 375–395, 2000.

[143] F. Tudisco and D. J. Higham, "Core-periphery detection in hypergraphs," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 1, pp. 1–21, 2023.

[144] A. Ceria and H. Wang, "Temporal-topological properties of higher-order evolving networks," *Scientific Reports*, vol. 13, no. 1, p. 5885, 2023.

[145] L. Gallo, L. Lacasa, V. Latora, and F. Battiston, "Higher-order correlations reveal complex memory in temporal hypergraphs," *arXiv preprint arXiv:2303.09316*, 2023.

[146] I. Iacopini, M. Karsai, and A. Barrat, "The temporal dynamics of group interactions in higher-order social networks," *arXiv preprint arXiv:2306.09967*, 2023.

[147] L. Di Gaetano, F. Battiston, and M. Starnini, "Percolation and topological properties of temporal higher-order networks," *Physical Review Letters*, vol. 132, no. 3, p. 037401, 2024.

[148] M. Mancastroppa, I. Iacopini, G. Petri, and A. Barrat, "The structural evolution of temporal hypergraphs through the lens of hyper-cores," *arXiv preprint arXiv:2402.06485*, 2024.

[149] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[150] G. Csardi, T. Nepusz, *et al.*, "The igraph software package for complex network research," *InterJournal, complex systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[151] T. P. Peixoto, "The graph-tool python library," *figshare*, 2014.

[152] A. Antelmi, G. Cordasco, B. Kamiński, P. Prałat, V. Scarano, C. Spagnuolo, and P. Szufel, "Analyzing, exploring, and visualizing complex networks via hypergraphs using simplehypergraphs. jl," *arXiv preprint arXiv:2002.04654*, 2020.

[153] D. J. Marchette, "HyperG," 2021.

[154] "HyperNetworkX," 2021.

[155] L. P. Diaz and M. P. Stumpf, "Hypergraphs. jl: representing higher-order relationships in julia," *Bioinformatics*, vol. 38, no. 14, pp. 3660–3661, 2022.

[156] A. Badie-Modiri and M. Kivelä, "Reticula: A temporal network and hypergraph analysis software package," 2022.

[157] N. Landry, L. Torres, I. Iacopini, M. Lucas, G. Petri, A. Patania, and A. Schwarze, "XGI," 2022.

[158] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." `http://snap.stanford.edu/data`, June 2014.

[159] R. A. Rossi and N. K. Ahmed, "An interactive data repository with visual analytics," *SIGKDD Explor.*, vol. 17, no. 2, pp. 37–41, 2016.

[160] T. P. Peixoto, "The netzschleuder network catalogue and repository," Apr. 2023.

[161] C. Berge, *Graphs and hypergraphs*. North-Holland Pub. Co., 1973.

[162] L. Torres, A. S. Blevins, D. Bassett, and T. Eliassi-Rad, "The why, how, and when of representations for complex systems," *SIAM Review*, vol. 63, no. 3, pp. 435–485, 2021.

[163] S. G. Aksoy, C. Joslyn, C. O. Marrero, B. Praggastis, and E. Purvine, "Hypernetwork science via high-order hypergraph walks," *EPJ Data Science*, vol. 9, no. 1, p. 16, 2020.

[164] Q. F. Lotito, F. Musciotto, F. Battiston, and A. Montresor, "Exact and sampling methods for mining higher-order motifs in large hypergraphs," *arXiv preprint arXiv:2209.10241*, 2022.

[165] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[166] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, 2006.

[167] N. Ruggeri, M. Contisciani, F. Battiston, and C. De Bacco, "Generalized inference of mesoscale structures in higher-order networks," *In preparation*, 2022.

[168] S. Miccichè and R. N. Mantegna, "A primer on statistically validated networks," *Computational Social Science and Complex Systems*, vol. 203, p. 91, 2019.

[169] F. Musciotto, F. Battiston, and R. N. Mantegna, "Identifying maximal sets of significantly interacting nodes in higher-order networks," *arXiv preprint arXiv:2209.12712*, 2022.

[170] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, 2004.

[171] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, vol. 78, no. 4, p. 046110, 2008.

[172] N. Ruggeri, F. Battiston, and C. De Bacco, "A principled, flexible and efficient framework for hypergraph benchmarking," *arXiv preprint arXiv:2212.08593*, 2022.

[173] G. Petri and A. Barrat, "Simplicial activity driven model," *Physical review letters*, vol. 121, no. 22, p. 228301, 2018.

[174] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge university press, 2008.

[175] B. C. Coutinho, A.-K. Wu, H.-J. Zhou, and Y.-Y. Liu, "Covering problems and core percolations on hypergraphs," *Physical Review Letters*, vol. 124, no. 24, p. 248301, 2020.

[176] T. Carletti, D. Fanelli, and S. Nicoletti, "Dynamical systems on hypergraphs.," *arXiv:2006.01243*, 2020.

[177] R. Muolo, L. Gallo, V. Latora, M. Frasca, and T. Carletti, "Turing patterns in systems with high-order interactions," *Chaos, Solitons & Fractals*, vol. 166, p. 112912, 2023.

[178] Y. Zhang, M. Lucas, and F. Battiston, "Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes," *Nature Communications*, vol. 14, no. 1, p. 1605, 2023.

[179] G. F. de Arruda, G. Petri, and Y. Moreno, "Social contagion models on hypergraphs," *Phys Rev Res*, vol. 2, no. 2, p. 023032, 2020.

[180] G. St-Onge, H. Sun, A. Allard, L. Hébert-Dufresne, and G. Bianconi, "Universal nonlinear infection kernel from heterogeneous exposure on higher-order networks," *Physical review letters*, vol. 127, no. 15, p. 158301, 2021.

[181] A. Civilini, N. Anbarci, and V. Latora, "Evolutionary game model of group choice dilemmas on hypergraphs," *Physical Review Letters*, vol. 127, no. 26, p. 268301, 2021.

[182] A. Civilini, O. Sadekar, F. Battiston, J. Gómez-Gardenes, and V. Latora, "Explosive cooperation in social dilemmas on higher-order networks," *arXiv preprint arXiv:2303.11475*, 2023.

[183] N. W. Landry, M. Lucas, I. Iacopini, G. Petri, A. Schwarze, A. Patania, and L. Torres, "Xgi: A python package for higher-order interaction networks," *Journal of Open Source Software*, vol. 8, no. 85, p. 5162, 2023.

[184] A. Antelmi, D. De Vinco, and C. Spagnuolo, "Hypergraphrepository: A community-

driven and interactive hypernetwork data collection," in *International Workshop on Algorithms and Models for the Web-Graph*, pp. 159–173, Springer, 2024.

[185] `https://journals.aps.org/datasets`, 2021.

[186] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[187] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 555–564, ACM, 2017.

[188] "SocioPatterns collaboration," *http://www.sociopatterns.org/*, 2008.

[189] L. Jure, "Snap datasets: Stanford large network dataset collection," *Retrieved December 2021 from http://snap. stanford. edu/data*, 2014.

[190] J. Wu, J. Liu, W. Chen, H. Huang, Z. Zheng, and Y. Zhang, "Detecting mixing services via mining bitcoin transaction network with hybrid motifs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 4, pp. 2237–2249, 2021.

[191] N. Yadati, V. Nitin, M. Nimishakavi, P. Yadav, A. Louis, and P. Talukdar, "Nhp: Neural hypergraph link prediction," in *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1705–1714, 2020.

[192] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*, pp. 243–246, 2015.

[193] N. Yadati, T. Gao, S. Asoodeh, P. Talukdar, and A. Louis, "Graph neural networks for soft semi-supervised learning on hypergraphs," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 447–458, Springer, 2021.

A | Appendix A: Additional dataset descriptions

A.1 Undirected hypergraphs

The datasets gathered to perform our motif analysis of real-world higher-order systems come from a variety of domains: sociology (proximity contacts, votes), technology (e-mails), biology (gene/disease, drugs) and co-authorship.

- **Gene/disease**: Nodes correspond to genes and each hyperedge is the set of genes associated to a disease.
- **NDC_classes**: Nodes correspond to classification labels and each hyperedge is the set of all the labels applied to a drug. Each drug is timestamped with the day it was first marketed.
- **NDC_substances**: Nodes correspond to substances and each hyperedge is the set of substances in a drug. Each drug is timestamped with the day it was first marketed.
- **DBLP**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication tracked by DBLP. Each paper is timestamped with the year of publication.
- **History**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of history. Each paper is timestamped with the year of publication.
- **Geology**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of geology. Each paper is timestamped with the year of publication.
- **PACS**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of physics. Each paper is timestamped with the year of publication and with its PACS classification. The latter allows to split the set of papers in ten subfields of physics by using the highest hierarchical level of PACS classification.
- **email-EU**: Nodes correspond to users in a European research institution and each hyperedge consists of the sender and all recipients of an email. Each email is timestamped.
- **email-ENRON**: Nodes correspond to Enron employees and each hyperedge consists of the sender and all recipients of an email. Each email is timestamped.
- **Wiki**: Nodes correspond to Wikipedia users and each hyperedge is the set of users

that expressed the same vote in a Wikipedia voting event since the inception of Wikipedia until January 2008.

- **Justice**: Nodes correspond to justices of the Supreme Court in the US and each hyperedge is the set of justices that expressed the same vote in a case from 1946 to 2019.
- **Primary school**: Nodes are students of a primary school. Wearable sensors are exploited to construct a network of proximity contacts among the students. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **High school**: Nodes are students of a high school. Wearable sensors are exploited to construct a network of proximity contacts among the students. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Conference**: Nodes are participants of a conference. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Hospital**: Nodes are people at a hospital. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Workplace**: Nodes are employees of a company. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Baboons**: Nodes are baboons in an enclosure of a Primate Center in France. Wearable sensors are exploited to construct a network of proximity contacts among baboons. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.

The summary statistics of the datasets are reported in Supplementary Table 1.

Dataset	N	E	E_2	E_3	E_4	E_5	Domain
Gene/disease	9703	11181	1311	614	443	363	Bio
NDC_classes	1161	1088	297	121	125	94	Bio
NDC_substances	5311	9906	1130	745	535	500	Bio
DBLP	1924991	2466799	693363	667291	419434	205970	Co-auth
History	1014734	895439	160885	47423	19120	8775	Co-auth
Geology	1256385	1203895	275736	227950	159509	99140	Co-auth
PACS0	98478	75985	12168	8373	14068	3064	Co-auth
PACS1	67055	43957	10532	7869	4647	2015	Co-auth
PACS2	47475	27504	5424	3452	2224	1411	Co-auth
PACS3	33479	16977	2099	2105	2590	1169	Co-auth
PACS4	57533	41133	8977	6571	5139	2897	Co-auth
PACS5	14455	5306	1062	1041	719	477	Co-auth
PACS6	71989	41663	5140	3293	4946	1715	Co-auth
PACS7	82964	55227	9664	9325	8294	4915	Co-auth
PACS8	13451	5618	1659	1447	910	485	Co-auth
PACS9	18666	7515	2361	2143	1034	397	Co-auth
email-EU	998	25027	12753	4938	2294	1359	Tech
email-ENRON	143	1512	809	317	138	63	Tech
Wiki	6210	5925	593	427	338	304	Socio
Justice	38	2864	216	456	506	560	Socio
Primary school	242	12704	7748	4600	347	9	Socio
High school	327	7818	5498	2091	222	7	Socio
Conference	403	10541	8268	1861	258	63	Socio
Hospital	75	1825	1108	657	58	2	Socio
Workplace	92	788	742	44	2	0	Socio
Baboons	13	231	78	142	11	0	Socio

Table A.1: Details of the real-world networked datasets considered for our experiments. Each higher-order network is described by its number of nodes, its total number of hyperedges and its number of hyperedges of size 2, 3, 4 and 5.

A.2 Directed hypergraphs

This section provides detailed descriptions of the datasets used in our experiments. The datasets are originally collected in [128] and represent a diverse range of real-world systems with directed higher-order interactions. Summary statistics of the datasets used in our experiments are reported in Table A.2.

- **Question answering data.** We use two QNA datasets: Math-overflow and Server-fault, both sourced from Stack Exchange logs. A hyperedge $e_i = (S_i, T_i)$ indicates a question posted by the user in the target set T_i and answered by the users in the source set S_i. Each hyperedge has a unit target set, i.e., $|T_i| = 1, \forall i = \{1, ..., |E|\}$.
- **Email data.** We use two email datasets: email-enron [34] and email-eu [189]. A hyperedge $e_i = (S_i, T_i)$ represents an email where the sender is the source set S_i, and the receivers (including cc-ed users) form the target set T_i. Each hyperedge has a unit source set, i.e., $|S_i| = 1, \forall i = \{1, ..., |E|\}$.
- **Bitcoin transactions data.** We use three bitcoin transaction datasets: bitcoin-2014, bitcoin-2015, and bitcoin-2016 [190]. They contain the first 1 500 000 transactions in 11/2014, 06/2015, and 01/2016 respectively. A hyperedge $e_i = (S_i, T_i)$ corresponds to a transaction where the accounts from which the coins are sent form the source set S_i, and the accounts receiving the coins make up the target set T_i.
- **Metabolic data.** We use two methabolic datasets: iAF1260b and iJO1366 [191]. Nodes are the genes and hyperedges are metabolic reactions. A hyperedge $e_i = (S_i, T_i)$ indicates that the reaction among genes in the source set S_i results in genes in the target set T_i.
- **Citation data.** We use two citation datasets: citation-data mining and citation-software [192, 193]. A hyperedge $e_i = (S_i, T_i)$ represents a citation from a paper co-authored by the authors in the source set S_i to a paper co-authored by the authors in the target set T_i. Papers with more than 10 authors are filtered out.

Dataset	$\lvert V\rvert$	$\lvert E\rvert$	$\overline{\lvert S_i\rvert}$	$\overline{\lvert T_i\rvert}$
bitcoin-2014	1 697 625	1 437 082	1.478	1.697
bitcoin-2015	1 961 886	1 449 827	1.568	1.744
bitcoin-2016	2 009 978	1 451 135	1.495	1.715
metabolic-iaf1260b	1 668	2 083	1.998	2.267
metabolic-iJO1366	1 805	2 251	2.026	2.272
email-enron	110	1 484	1.000	2.354
email-eu	986	35 772	1.000	2.368
citation-dm	27 164	73 113	3.253	3.038
citation-software	16 555	53 177	2.927	2.717
qna-math	34 812	93 731	1.779	1.000
qna-server	172 330	272 116	1.747	1.000

Table A.2: Summary statistics of the datasets used in our experiments.

B | Appendix B: Additional analyses

B.1 More on motif analysis in hypergraphs

B.1.1 Building hyperedges from dyadic data

While in some datasets higher-order structures are naturally encoded as hyperedges (e.g. three authors collaborating on the same paper), in others (e.g. face-to-face interactions) one needs to infer higher-order structures from dyadic interactions. The simplest way to perform this task is to set time-windows of size t, build the aggregated network for each time-window, and promote cliques of size n to hyperedges of size n. The idea is that if every node in a set of n nodes is interacting with every other node in the set, then the original interaction was probably a group interaction.

The choice of the size of the time window could be an important factor when evaluating the experimental results. The smallest timescale available in most of the datasets is 20 seconds, which represents our originally selected time window. In Figure B.1, we report the change in the distribution of the occurrences of the higher-order motifs of order 3 with changes in the considered time window. We considered the dataset High School and time windows of 20, 40, 60, 80, 100 and 120 seconds. At these scales, the choice of the time window does not seem to take a role in the results. Please notice that larger time windows would badly approximate the notion of *group interaction*, since they would aggregate pairwise interactions happened far apart in time.

B.1.2 Clustering hypergraphs via their 4-motifs profiles

Higher-order motifs of order 4 are able to better capture structural information compared to their counterparts of order 3. Intuitively, one can notice a richer hierarchical intra-cluster organization from the dendrograms on the left of the correlation matrices (Fig. 2b and 3b). Moreover, a better separation between the two clusters can be noticed from the lowering of the mean correlation between the SPs of hypergraphs belonging to different clusters (practically, we can notice less red squares and more blue squares that separate the clusters).

More formally, we can consider the scatter plot in Supplementary Figure B.2. Each point $p = (x, y)$ is described by $x = corr_3(d_1, d_2)$ and $y = corr_4(d_1, d_2)$ where $corr_n(d_1, d_2)$ is the correlation between the SPs of the higher-order motifs of order n of the datasets d_1 and d_2. Every possible pair of datasets is considered. In red we plot points in which the two datasets belong to the same big cluster (e.g. both datasets in Bio/Co-auth), in blue we plot points in which the two datasets belong to different clusters (i.e. one in Bio/Co-auth and the other in Socio/Tech). We can observe a significant drop in the correlations of the

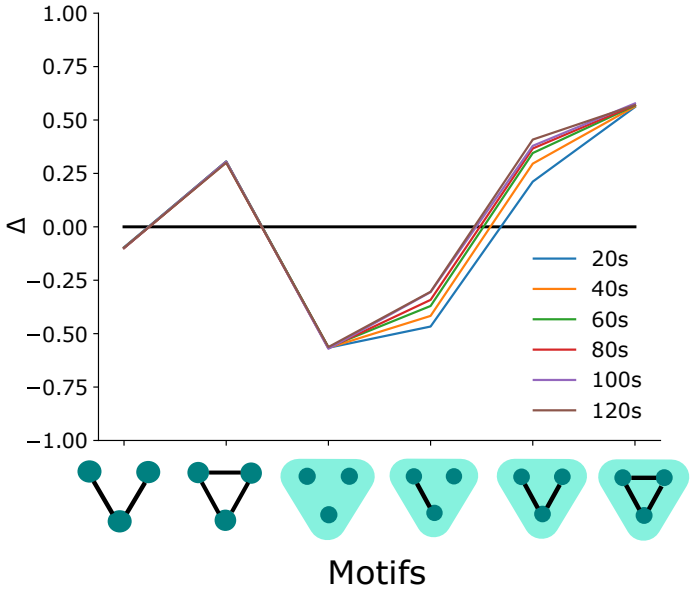

Figure B.1: Effect of the choice of the time window for promoting cliques to group interactions on the SP of the dataset High School. Each time window is expressed in seconds. Δ is the relative abundance of each motif with respect to the null model.

SPs of datasets from different clusters when considering motifs of order 4.

Taken together, these results confirm that scaling up the analysis of higher-order motifs to larger orders informs our tool with more information, capturing more nuanced and fine-grained wiring patterns.

B.1.3 Disaggregated 3-Motifs

In Figure B.3 we compute and show all the SPs of the real-world hypergraphs considered in our experiments, subdivided by their domain. In the main paper, we group and average them to extract a fingerprint of a domain. We can notice that hypergraphs from the same domain tend to have very correlated SPs, while hypergraphs from different domains do not display such correlation.

B.1.4 Most significative 4-motifs

In Figure B.4 we report the most under-expressed higher-order motifs of order 4 for each domain. Comparing Supplementary Figure B.4 with Fig.3c from the main paper (which shows the most over-expressed higher-order motifs of order 4 for each domain) further supports the idea that hypergraphs from the Socio / Tech domain have a preference to-

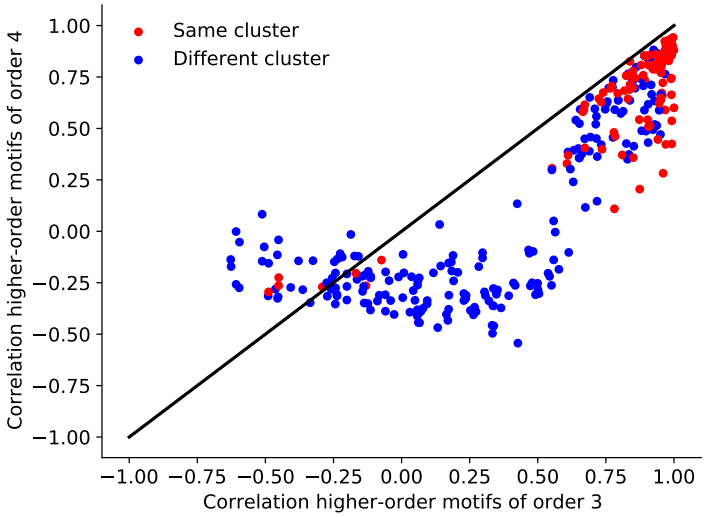

Figure B.2: Each point $p = (x, y)$ is described by $x = corr_3(d_1, d_2)$ and $y = corr_4(d_1, d_2)$ where $corr_n(d_1, d_2)$ is the correlation between the SPs of the higher-order motifs of order n of the datasets d_1 and d_2. Every possible pair of datasets is considered. In red we plot points in which the two datasets belong to the same big cluster (e.g. both datasets in Bio/Co-auth), in blue we plot points in which the two datasets belong to different clusters (i.e. one in Bio/Co-auth and the other in Socio/Tech).

wards over-expressing structures involving more lower-order nested relations (e.g. dyadic links), while hypergraphs from the Bio / Co-author domain display a preference towards less nested relations but of higher-order. For example, people interacting in groups are likely to interact also in single pairs, therefore it is unlikely that group interactions in the Socio / Tech domain are not supported by a large number of nested lower-level interactions. In fact, from Supplementary Figure B.4, we can notice that patterns involving a group interaction with none or few nested dyadic interactions are penalized. To name another example, representative of the opposite cluster, people tend to write papers in large groups and tend to maintain the same research group over time, with few additions or removals; therefore patterns involving only dyadic relations are penalized.

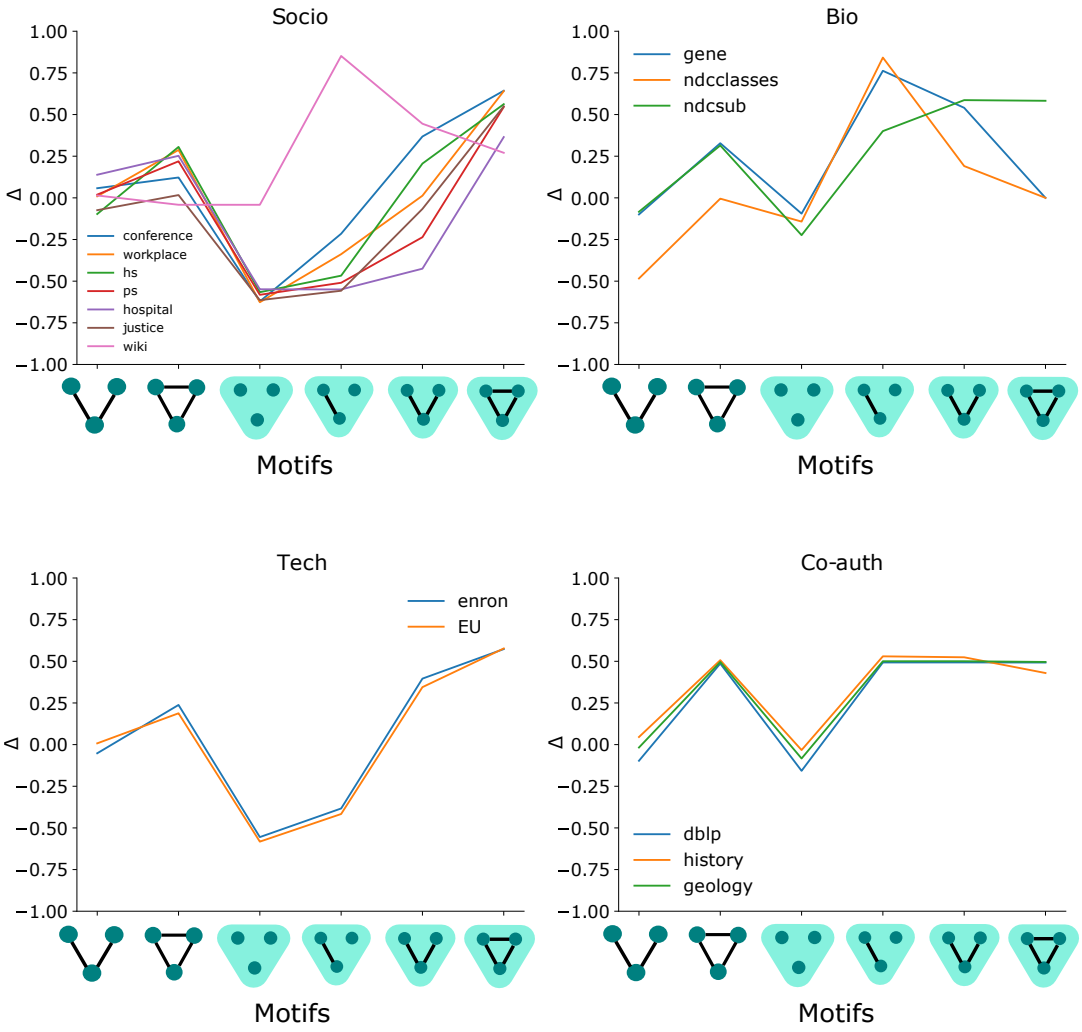

Figure B.3: All the SPs of the real-world hypergraphs considered in our experiments, subdivided by their domain. Hypergraphs from the same domain tend to have very correlated SPs, while hypergraphs from different domains do not display such correlation.

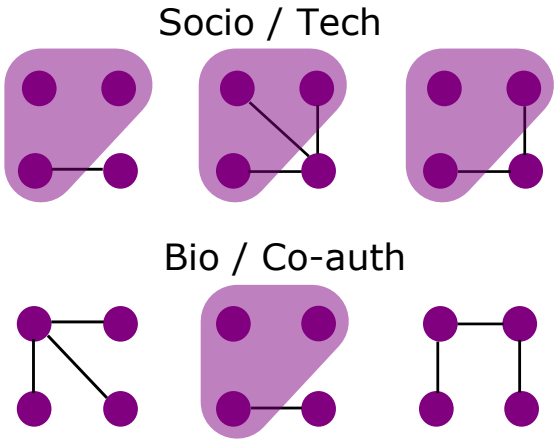

Figure B.4: Most under-expressed higher-order motifs of order 4 from the two clusters.

B.2 Multiplex measures on randomized hypergraphs

In this section, we investigate how our proposed multiplex measures for higher-order networks behave under re-wiring. This analysis helps in understanding the robustness of the observed network properties and the significance of the original structure.

We propose two different randomization methods. They are available as part of Hypergraphx [3]. Results have been averaged over ten realizations of the random models.

For this analysis we employ:

- A *configuration model* for hypergraphs [39] which preserves the degree sequences and the hyperedge order distributions of the original layers of the multiplex hypergraphs while randomizing the actual nodes involved in the hyperedges.
- A *random model* for hypergraphs which preserves only the hyperedge order distributions of the original layers of the multiplex hypergraphs while randomizing the actual nodes involved in the hyperedges.

By preserving the degree sequence in each layer, the configuration model maintains the node activity and the degree correlations across layers. Additionally, it preserves the distribution of overlapping degrees across different layers, and consequently, the node participation coefficient. However, by re-wiring interactions and thus randomizing the neighbors in each interaction, we disrupt the micro and mesoscale structure of the hypergraphs. To assess this, in Fig. B.5 we show the impact of the configuration model randomization process on the measure of hyperedge overlap in our empirical datasets. We notice a consistent drop in the number of overlapping hyperedges with respect to the original hypergraphs. In particular, after randomization, there are no hyperedges that appear in more than one layer in the case of movie collaborations.

Additionally, in Fig. B.6 we show that the correlations between community structures are not preserved, as nodes tend to lose their preferential patterns of interactions under randomization. In particular, in HIGH SCHOOL the algorithm collapses every node into a single community in every layer since the networks lack clear separations.

Our second random model is more disruptive, as it disintegrates structures to a greater extent. While this model preserves the hyperedge size distributions for each layer of a multiplex hypergraph, it randomizes the nodes involved in the interactions. This process is sufficient to dismantle node activity patterns and disrupt node degree correlations. Since node degrees are not preserved, node and hyperedge participation coefficients are also not

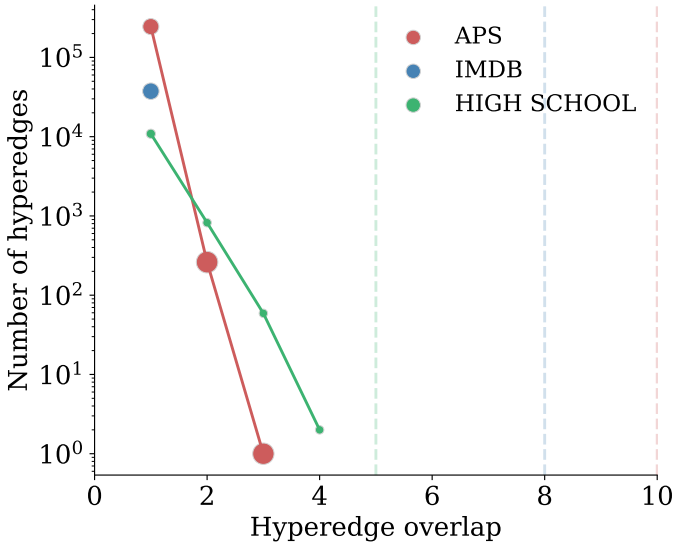

Figure B.5: Impact of configuration model randomizations on hyperedge overlap. We notice a consistent decrease in overlapping hyperedges compared to the original data, with no hyperedges appearing in more than one layer for movie collaborations after randomization.

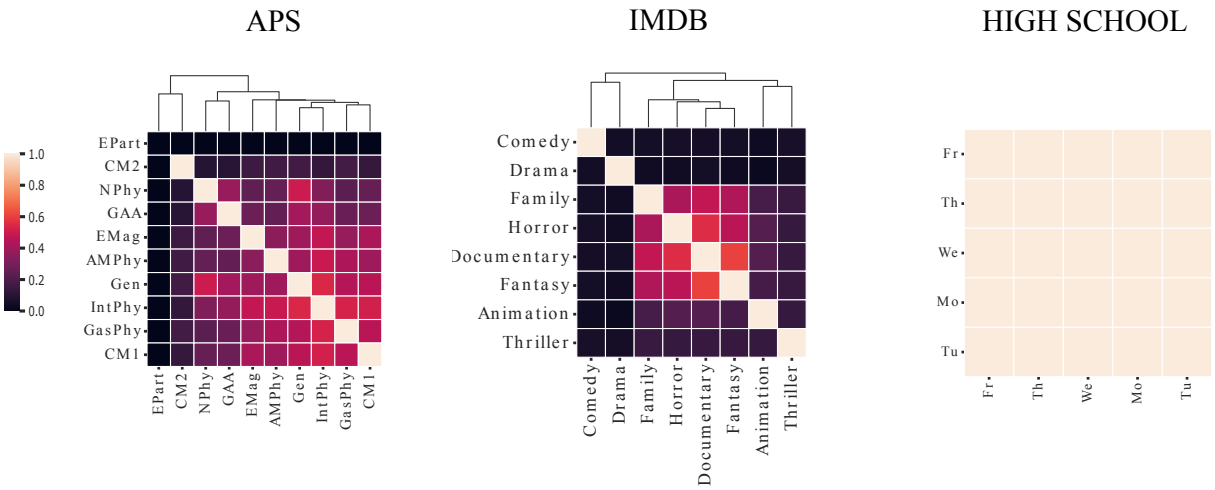

Figure B.6: Configuration model randomization leads to the loss of preferential interaction patterns, with nodes losing correlated community affiliations across layers.

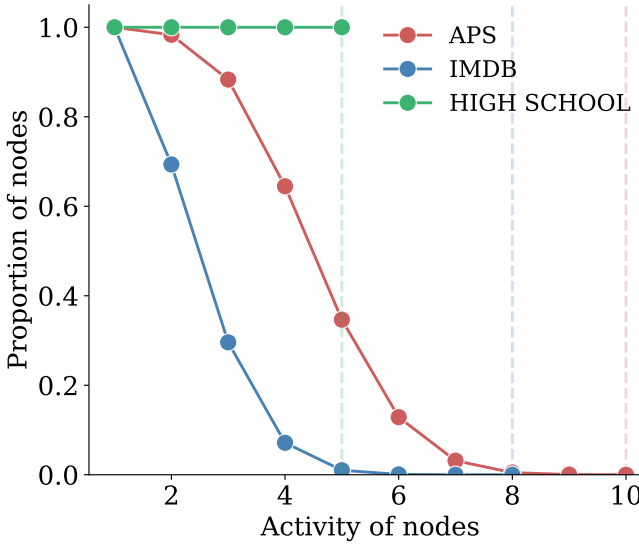

Figure B.7: Rewiring preserves layer density but increases node activity across multiple layers. This is shown by the increase in nodes' simultaneous activity across layers compared to the original data.

maintained. It is important to notice that since all layers of a multiplex hypergraph share the same node set, the only property of a layer that is preserved is layer density. The rewiring of interactions allows more nodes to become active in new layers. This is shown in Fig. B.7, in which we analyze nodes' simultaneous activity across multiple layers and see an increase in such statistics with respect to original data.

The disruption of activity patterns is evident when computing the similarity among activity matrices between layers (Fig. B.8). In fact, the cosine similarity of the activity matrices drops to zero for both APS and IMDB, while HIGH SCHOOL maintains a generally high level of similarity (explained by the abundance of interactions of size two and the high level of activity across layers). The disruption of activity patterns in APS and IMDB is caused by the rewiring of hyperedges with large cardinality, especially when these large hyperedges overlap in real-world data.

Finally, it is important to examine the degree correlations across layers and the participation coefficient of nodes. The rewiring process significantly affects degree correlations, resulting in all pairs of layers being uncorrelated in terms of node degrees (Fig. B.9a). In Fig. B.9b we show that hubs tend to disappear (lower overlapping degrees than original data), average interaction size tends to decrease (due to overabundance of low order interaction in the data) and participation coefficient tends to increase (given by the increase

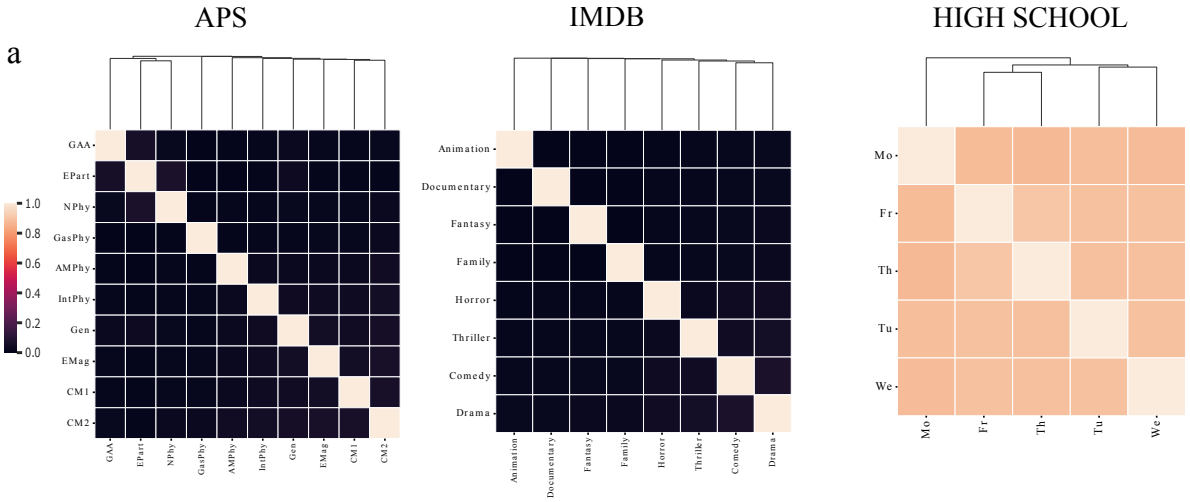

Figure B.8: Similarity among activity matrices between layers drops to zero for APS and IMDB but remains high for HIGH SCHOOL. This is due to the rewiring of large cardinality hyperedges, particularly when these hyperedges overlap in real-world data.

in activity of nodes).

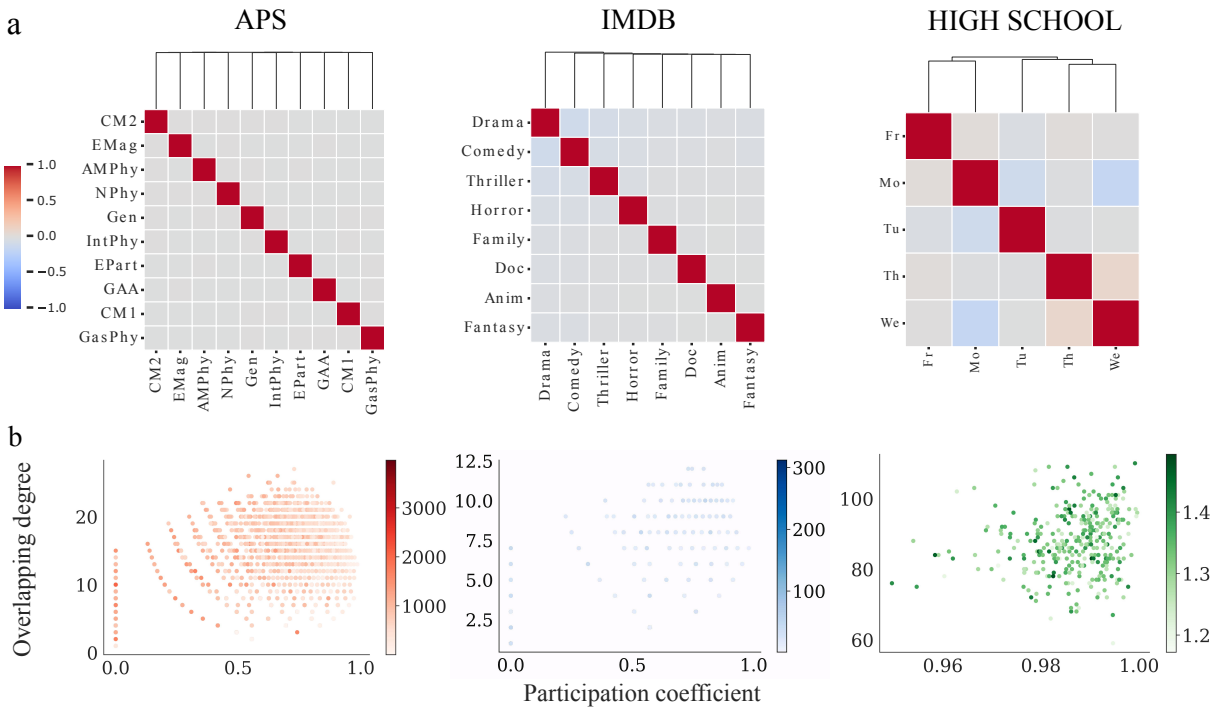

Figure B.9: Effects of rewiring on multiplex hypergraphs properties. (a) Rewiring results in uncorrelated node degrees across all layer pairs. (b) Hubs disappear (lower overlapping degrees), average interaction size decreases, and participation coefficient increases due to higher node activity.

C | Appendix C: Additional algorithmic details

C.1 Parameters search in approximated motif mining algorithms

Our approximated algorithm requires different parameters. The parameter S controls the number of samples of hyperedges to be performed to estimate the count of the patterns of sub-hypergraphs. Without a careful design, directly sampling hyperedges from the hypergraphs leads to unreliable results. In fact, the distribution of the size of the hyperedges in a real-world hypergraph is not uniform, leading the algorithm to often sample hyperedges of size 2, and rarely, for example, hyperedges of size 4. This would result in poor estimations of higher-order motifs involving a group interaction of size 3 or 4. To address this issue, we stratify our sampling process, allocating specific sample budgets to hyperedges of different sizes. This ensures a balanced representation of hyperedges across all sizes. Let S_k be the number of samples assigned to hyperedges of size k. We fix the sum of S_k for every k to be equal to S. We estimate empirically good values for the parameters S_k, exhaustively searching among different combinations of values and selecting those that maximize a defined quality function (Pearson's correlation ρ between the exact higher-order motif profile and the estimated one). We perform the analyses on two datasets, one for each macro-domain: `high school` and `history`. We consider motifs of order 4, therefore we need to estimate S_2, S_3 and S_4, namely, respectively the number of samples from the hyperedges of order 2, 3 and 4. Given that $S_2 + S_3 + S_4 = S$, one can fix S_2, parametrize S_3 and S_4 to be multiple of S_2, and perform exhaustive search. We show the results in Figure C.1. Averaging the results of the two matrices, we get that our quality measure is maximized when $S_3 = 3S_2$ and $S_4 = 2S_2$. We use these parameters in our experiments.

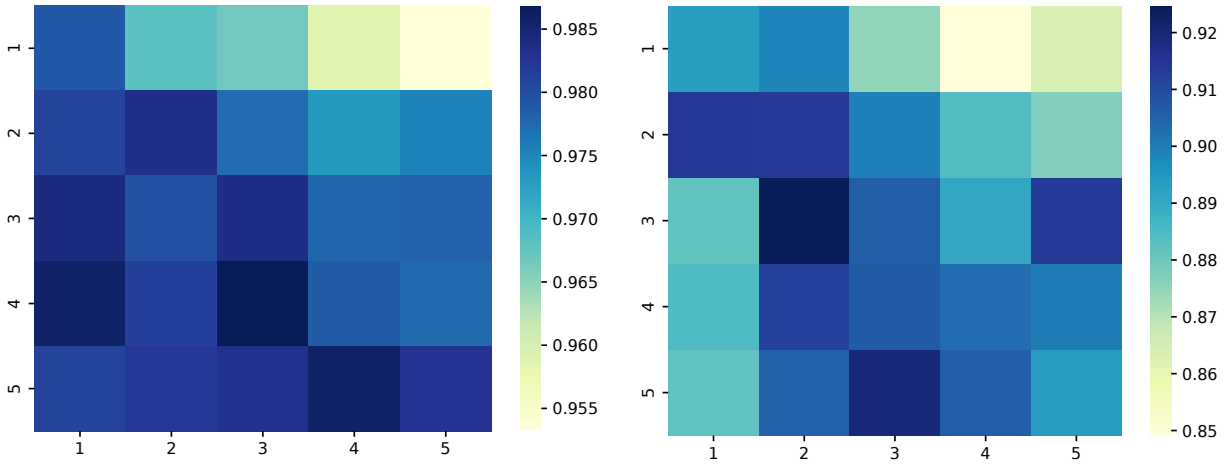

Figure C.1: We parametrize the number of samples of hyperedges of size 3 and 4 with respect to the number of samples of hyperedges of size 2 and search the values for which the correlation between the exact motif profile and the estimated one is maximized. The x-axis parametrizes the number of samples of hyperedges of size 4. The y-axis parametrizes the number of samples of hyperedges of size 3. Light squares exhibit lower levels of correlation, while dark squares show higher levels. On the left, we show the matrix for the high school dataset. On the right, is the matrix for the history dataset. We get the best parameters by averaging the two matrices.

C.2 Computing reciprocity in directed hypergraphs

Here, we outline the algorithms for measuring reciprocity in directed hypergraphs efficiently.

- **Exact reciprocity.** Each hyperedge $e = (s, t)$ is stored in a hash-based dictionary, and for each hyperedge, we search for a reverse hyperedge $e' = (t, s)$. Since each lookup takes constant time, the overall complexity is $O(m)$, where m is the number of hyperedges.
- **Strong reciprocity.** For each hyperedge $e = (s, t)$, we maintain a reachability dictionary that tracks which nodes in the target set t can reach other nodes via multiple hyperedges. We then check whether the source set s is fully covered by the accumulated reachable nodes from the target set t. This involves iterating over each hyperedge, for each target node, accumulating the reachable nodes and then checking if the source set is a subset of this accumulated set. Computing the union of reachable nodes is $O(s \cdot t)$, where s is the maximum size of source sets and t is the maximum size of target sets. This operation is repeated for all hyperedges, leading to a total complexity of $O(m \cdot s \cdot t)$.
- **Weak reciprocity.** First, we construct a dictionary to store all directed node pairs between the source and target sets of each hyperedge. Then, for each hyperedge, we check whether any of its target nodes are linked back to the source nodes via reverse connections in the dictionary. The computational complexity is dominated by the first operation, which is $O(m \cdot s \cdot t)$, where s is the maximum size of the source sets and t is the maximum size of the target sets across all hyperedges.

In practice, executing these algorithms on the real-world datasets used in our experiments requires only a few minutes for all datasets combined, demonstrating the computational efficiency of the proposed methods.

C.3 Algorithms for motif analysis in directed hypergraphs

In order to design efficient algorithms for mining directed higher-order motifs, we extend prior ideas developed for the same problem in undirected hypergraphs [2]. Our algorithms are efficient enough to count motifs of size 3 and 4 in datasets of reasonable size (comparable to those used in our experiments). In Fig. C.2, we show the execution times of our algorithms for motifs of order 3 and 4 across various datasets, highlighting the increase in

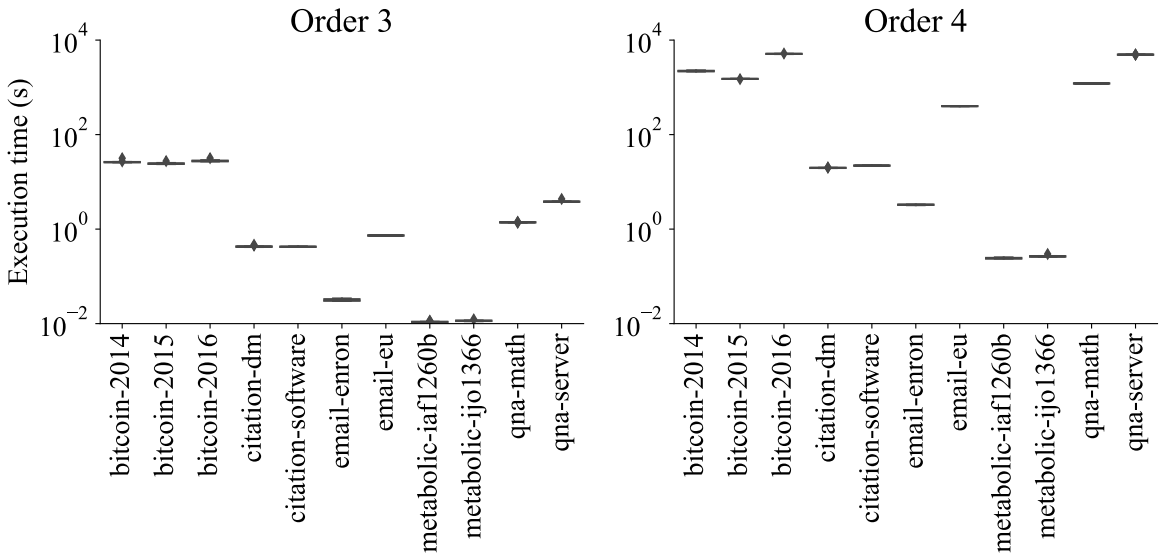

Figure C.2: Execution times in seconds of the algorithms for mining motifs of order 3 and 4 across datasets. We consider 10 trials for each dataset.

time when moving from order 3 to order 4. Scaling to larger datasets and motifs of larger size would require more sophisticated approaches, such as sampling algorithms [2], which we leave for future work.

The algorithm for mining motifs (involving at least one group interaction) of order 3 begins by iterating through each hyperedge in the hypergraph that contains exactly three vertices. For each such hyperedge, it identifies all possible subsets of vertices and checks whether one or more subsets form valid directed hyperedges in the hypergraph. Valid subsets, along with the original hyperedge, define the motif structure involving those three vertices. To ensure consistency in motif identification, the algorithm generates a canonical form of the motif by lexicographical ordering its vertices and edges, which can be computed by sorting the $n!$ possible relabels. This canonical representation allows motifs with the same structural pattern to be compared and counted, even if they differ in their vertex labels. Each canonical form of motifs is stored in a frequency hash map. If the motif has not been encountered before, it is added to the map; if it has, its frequency count is incremented. In the end, the algorithm outputs a distribution of the various motif structures of order 3. This algorithm operates in linear time with respect to the number of hyperedges of order 3. Specifically, its computational complexity is $O(m_3)$, where m_3 is the number of hyperedges involving exactly three vertices. Each motif construction and comparison is performed in constant time due to the fixed size of the motifs.

The algorithm for mining motifs of order 4 follows a similar approach. First, it iterates over

all hyperedges of size 4, counting the motifs involving exactly these 4 nodes. Unlike the previous algorithm, it then iterates over all hyperedges of size 3, performing an additional neighborhood exploration step to identify the fourth node involved in the motif. Each neighboring node is considered during this process. Once the 4 nodes are identified, the algorithm constructs the motif as before.

In Algorithm C.1 and Algorithm C.2, we present detailed pseudocode for the algorithms designed to count directed higher-order motifs of sizes 3 and 4, respectively.

Algorithm C.1 Motifs of order 3

Input: A directed hypergraph $\mathcal{H} = (V, E)$

Output: distribution of the frequency of the motifs of order 3

Let M be the motifs frequency hash map
Let U be the isomorphism class hash map
for each hyperedge e of order 3 in E **do**
    $V^* \leftarrow$ vertices of e
    $motif \leftarrow \emptyset$
    for each $e^* \in \mathcal{P}(V^*)$ **do**
        if $e^* \in E$ **then**
            $motif \leftarrow motif \cup e^*$
        end if
    end for each
    $C_m \leftarrow$ lexicographically minimum canonic relabel of *motif*
    if $C_m \notin M$ **then**
        $M[C_m] \leftarrow 0$
    end if
    $M[C_m] += 1$
    Set vertices of *motif* as visited
end for each

Algorithm C.2 Motifs of order 4

Input: A directed hypergraph $H = (V, E)$

Output: distribution of the frequency of the motifs of order 4

Let M be the motifs frequency hash map
Let s be the isomorphism class hash map
for each hyperedge e of order 4 in E **do**
    $V^* \leftarrow$ vertices of e
    $motif \leftarrow \emptyset$
    for each $e^* \in \mathcal{P}(V^*)$ **do**
        if $e^* \in E$ **then**
            $motif \leftarrow motif \cup e^*$
        end if
    end for each
    $C_m \leftarrow$ lexicographically minimum canonic relabel of *motif*
    if $C_m \notin M$ **then**
        $M[C_m] \leftarrow 0$
    end if
    $M[C_m] += 1$
    Set vertices of *motif* as visited
end for each
$\mathcal{H} \leftarrow$Discard all hyperedges of order 4 from $\mathcal{H}$
for each hyperedge e of order 3 in E **do**
    Let $\mathcal{Z}$ be the set of hyperedges adjacent to e
    for each hyperedge ζ in $\mathcal{Z}$ **do**
        if $|\zeta \cup e| = 4$ and $\zeta \cup e$ not already visited **then**
            $V^* \leftarrow$ vertices of $\zeta \cup e$
            $motif \leftarrow \emptyset$
            for each $e^* \in \mathcal{P}(V^*)$ **do**
                if $e^* \in E$ **then**
                $motif \leftarrow motif \cup e^*$
                end if
            end for each
            $C_m \leftarrow$ lexicographically minimum canonic relabel of *motif*
            if $C_m \notin M$ **then**
                $M[C_m] \leftarrow 0$

```
            end if
            M[C_m] += 1
            Set vertices of motif as visited
        end if
    end for each
end for each
```

List of Figures

List of Tables