Applied Network Science

RESEARCH Open Access

Annotated hypergraphs: models and applications

Philip Chodrow[1*†] and Andrew Mellor[2†]

*Correspondence:
pchodrow@mit.edu
†Philip Chodrow and Andrew Mellor contributed equally to this work.
[1]Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 02139 Cambridge, MA, USA
Full list of author information is available at the end of the article

Abstract
Hypergraphs offer a natural modeling language for studying polyadic interactions between sets of entities. Many polyadic interactions are asymmetric, with nodes playing distinctive roles. In an academic collaboration network, for example, the order of authors on a paper often reflects the nature of their contributions to the completed work. To model these networks, we introduce *annotated hypergraphs* as natural polyadic generalizations of directed graphs. Annotated hypergraphs form a highly general framework for incorporating metadata into polyadic graph models. To facilitate data analysis with annotated hypergraphs, we construct a role-aware configuration null model for these structures and prove an efficient Markov Chain Monte Carlo scheme for sampling from it. We proceed to formulate several metrics and algorithms for the analysis of annotated hypergraphs. Several of these, such as assortativity and modularity, naturally generalize dyadic counterparts. Other metrics, such as local role densities, are unique to the setting of annotated hypergraphs. We illustrate our techniques on six digital social networks, and present a detailed case-study of the Enron email data set.

Keywords: Hypergraphs, Null models, Network science, Statistical inference, Community detection

Introduction

Many data sets of contemporary interest log interactions between sets of entities of varying size. In collaborations between scholars, legislators, or actors, a single project may involve an arbitrary number of agents. A single email links at least one sender to one or more receivers. A given chemical reaction may require a large set of reagents. The dynamics of processes such as these may depend on these polyadic interactions, and in many cases cannot equivalently expressed through constituent pairwise interactions. This phenomenon is observed in areas as diverse as knowledge aggregation (Greening Jr et al. 2015), social contagion (de Arruda et al. 2019), and the evolution of cooperation (Tarnita et al. 2009), among many others. Because of this, these networks cannot be represented by via the classical paradigm of dyadic graphs without a significant loss of model fidelity.

Polyadic data representations such as hypergraphs (Berge 1984; Chodrow 2019a) and simplicial complexes (Young et al. 2017; Carlsson 2009) have therefore emerged as practical modeling frameworks that directly represent interactions between arbitrary sets of agents. As shown, for example, by one of the present authors in Chodrow (2019a), the choice of whether and when to represent polyadic data dyadically can lead to directionally contrasting study conclusions when studying common network metrics such as triadic closure and degree-assortativity. Furthermore, the use of polyadic data representations enable the analyst to study measures of higher-order structure which cannot even be defined in the dyadic framework.

In some cases, even polyadic data representations may be inadequate. Increasingly, network data sets incorporate rich metadata over and above topological structure. Models that flexibly incorporate this information can assist analysts in discovering features that may not be apparent without metadata. In this article, we consider an important and relatively general class of metadata in which nodes are assigned *roles* in each edge. A variety of social data sets involve such roles. For example, research articles have junior and senior authors. Political bills have sponsors and supporters. Movies have starring and supporting actors. Emails have senders, receivers, and carbon copies. Chemical reactions possess reactants, solvents, catalysts and inhibitors. These roles induce asymmetries in edges, and permuting role labels within an edge results in a meaningfully different data set. For example, a movie in which actor *A* plays a starring role and *B* a supporting role becomes a different movie if the roles are exchanged.

Metadata, including roles, can be especially important for modeling processes evolving on network substrates. A trivial example is that information cannot flow along an email edge from receiver to sender. A less trivial example comes from a recent study (Rotabi et al. 2017), which found that conventions in scholarly documentation preparation tend to flow along collaborations from more senior authors to more junior ones. In many fields, senior authors will tend to be "last" authors, while junior ones are more likely to be "first" or "middle" authors. The author order therefore carries important information about the spread of conventions along this collaboration network.

There exists an extensive literature studying graphs and hypergraphs with metadata attached to nodes (Ghoshal et al. 2009; McMorris et al. 1994; Kovanen et al. 2013; Henderson et al. 2012; Peel et al. 2017) and edges (Mucha et al. 2010; Gomez et al. 2013; Battiston et al. 2014). The problem of studying hypergraphs with general roles, however, does not neatly fit into any of these frameworks. This is because roles are not attributes of either nodes or edges, but rather of node-edge pairs. An actress is not (intrinsically) a "lead actress" – she may play a leading role in one film and a supporting role in the next. Contextual metadata is familiar in the context of directed networks. Each edge contains two nodes, one of which possesses the role "source" and the other "target," however a node may be the source of some edges, and the target of others. In Gallo et al. (1993); Gallo and Scutella (1998), the authors allow edges to contain arbitrary numbers of nodes, each of which is assigned one of these two roles. This results in *directed hypergraphs*, which have found some application in the study of cellular networks (Klamt et al. 2009) and routing (Marcotte and Nguyen 1998) problems. Subsequent work generalized further to "multimodal networks" by introducing a relationship of "association" (Heath and Sioson 2009) alongside the source and target roles.

Several extant papers explicitly model roles in interactions. An early pair of papers by Söderberg (Söderberg 2003a; 2003b) define and study a class of *inhomogeneous random graphs* in which nodes possess colored (role-labeled) stubs denoting their role in a given interaction. This class of models preserves degree distributions in expectation rather than deterministically, and the author develops them only for dyadic graphs. A later paper by

Karrer and Newman (2010) assigns nodes to roles in network motifs – small recurrent subgraphs – and uses these roles to construct configuration-like models. A model closely related to the one we develop here can be obtained by using labeled cliques as the relevant motifs and interpreting these cliques as hyperedges. Most recently, Allard et al. (2015) define a flexible generalization of stub-matching that can reproduce a wide variety of network topologies, including the presence of heterogeneous hyperedges. As we discuss below, stub-matching can often generate graphs with a small number of structural degeneracies. When studying dynamics on graphs as the authors do, these degeneracies can often be ignored. Since our focus is primarily inferential, we instead choose an approach that explicitly avoids degeneracies.

Our aim in this work is to develop a unified modeling and analysis framework for polyadic data with contextual roles, which can then be flexibly deployed in varied domains. The article is structured as follows. We first define *annotated hypergraphs*, which naturally generalize the notion of directedness to polyadic data. We then define a configuration model for annotated hypergraphs, and prove a Markov Chain Monte Carlo algorithm for sampling from this model. Next, we formulate a range of role-aware metrics for studying the structure of annotated hypergraphs. Some of these are direct generalizations of familiar tools, including centrality, assortativity, and modularity. Others, such as local role densities, are qualitatively novel. We then bring our methods to bear on a small collection of social network data sets, showing how the framework of annotated hypergraphs allows us to flexibly highlight interpretable features in the data. Additionally, we conduct an extended case-study of the popular Enron email data set. We conclude with a discussion of our results and suggestions for future work in the modeling of rich, polyadic data. Throughout our development, we emphasize how the incorporation of metadata allows the analyst to ask and answer a wide array of questions not accessible through the pure hypergraph formalism of nodes and edges.

Annotated hypergraphs

An annotated hypergraph a hypergraph accompanied by additional metadata.

Definition 1 (Annotated Hypergraph) *An annotated hypergraph* $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \ell)$ *consists of:*

(1) *A node set* $\mathcal{V}$*.*

(2) *A labeled edge set* $\mathcal{E}$*, a multiset of subsets of* $\mathcal{V}$*. In particular, multi-edges are permitted, but edges in which the same node appears twice are not.*

(3) *A finite label set* $\mathcal{X}$*.*

(4) *A role labeling function* $\ell : \{(v, e) \in \mathcal{V} \times \mathcal{E} | v \in e\} \to \mathcal{X}$*.*

The statement $\ell(v, e) = x$ is to be read as "node v has role x in edge e." We emphasize that the labeling function is contextual. Roles are assigned neither to nodes or to edges, but rather to node-edge pairs. There are two representations of annotated hypergraphs that will be useful in our subsequent development. Let $n = |\mathcal{V}|$, $m = |\mathcal{E}|$, and $p = |\mathcal{X}|$. Let $\mathbb{H}$ refer to the set of all annotated hypergraphs with n nodes, m hyperedges, and label alphabet $\mathcal{X}$.

Definition 2 (Labeled Incidence Array) *The labeled incidence array* $\mathbf{T} = \mathbf{T}(\mathcal{H}) \in \{0,1\}^{n \times m \times p}$ *of an annotated hypergraph is defined entrywise by*

$$t_{vex} = \begin{cases} 1 & v \in e \text{ and } \ell(v,e) = x \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that the labeled incidence array is distinct from the tensorial representation of directed hypergraphs in Xie and Qi (2016). An alternative representation is especially useful in the design of algorithms.

Definition 3 (Annotated Bipartite Graph) *The annotated bipartite graph* $\mathcal{B}(\mathcal{H}) = (\mathcal{V}', \mathcal{E}')$ *consists of*

- *A node set* $\mathcal{V}' = (\mathcal{V}, \mathcal{E})$
- *An edge set* $\mathcal{E}'$*, where an edge between v and e exists in* $\mathcal{E}'$ *iff* $v \in e$ *in* $\mathcal{E}$*. Each edge is labeled by* $\ell(v,e)$*, and may therefore be written* $(v, e; x)$*.*

Let $\mathbb{B}$ *be the set of annotated bipartite graphs.*

It is convenient to let $\mathcal{B}^x$ denote the annotated bipartite graph obtained by removing all edges from $\mathcal{B}$ except those with label x.

We generalize the notion of self-loops in graphs via the concept of *degeneracy*.

Definition 4 (Degeneracy) *An edge* $e \in \mathcal{E}$ *is role-degenerate if the same node v appears twice in e, with the same role. Edge e is simply degenerate if the same node appears twice in e, possibly in different roles. We call an annotated hypergraph* $\mathcal{H}$ *role-degenerate (resp. degenerate) if it contains any role-degenerate (resp. degenerate) edges.*

It is difficult to find a modeling justification for hypergraphs with role-degenerate edges, but degenerate edges may have modeling applications. For example, in email networks, it may be useful to log instances in which a sender also copies themsleves into the recipients list. In our experiments below, however, we work only on data with neither form of degeneracy. Throughout the remainder of the paper, we restrict $\mathbb{H}$ to the set of *nondegenerate* annotated hypergraphs.

A configuration model for annotated hypergraphs

We now define a configuration null model on nondegenerate, annotated hypergraphs. As has recently been emphasized by Fosdick et al. (2018), there are several distinct models that are often called "configuration models," and care is required in order to define and sample from them.

To do so, we define two sets of vectors that summarize the incidence array $\mathbf{T}$. For each x, define the vector $\mathbf{d}^x$ of nonnegative integers entrywise as

$$d_v^x = \sum_{e \in E} t_{vex} \,. \tag{2}$$

Similarly, define the vector $\mathbf{k}$ of nonnegative integers entrywise as

$$k_e^x = \sum_{v \in V} t_{vex} \,. \tag{3}$$

The vector $\mathbf{d}^x$ counts the number of times that each node plays role x in $\mathcal{H}$, while the vector $\mathbf{k}^x$ counts the number of nodes with role x in each edge. Both nodes of degree zero (which participate in no edges) and edges of dimension zero (which contain no nodes) are permitted. However, because we will later sample from a probability distribution via edge swaps, these nodes and edges play no role in sampling and can be ignored.

Let $\mathbf{D}$ be the matrix whose xth column is $\mathbf{d}^x$, and $\mathbf{K}$ the matrix whose xth column is $\mathbf{k}^x$. In a slight abuse of notation, we also regard $\mathbf{D}$ and $\mathbf{K}$ as functions of $\mathcal{H}$.

Definition 5 (Configuration Null Space) *The configuration null space* $\mathbb{C}_{\mathbf{D},\mathbf{K}} \subset \mathbb{H}$ *induced by degree-role matrix* $\mathbf{D}$ *and dimension-role matrix* $\mathbf{K}$ *is*

$$\mathbb{C}_{\mathbf{D},\mathbf{K}} = \{\mathcal{H} \in \mathbb{H} \ : \ \mathbf{D}(\mathcal{H}) = \mathbf{D} \ , \ \mathbf{K}(\mathcal{H}) = \mathbf{K}\} \ .$$

If $\mathbb{C}_{\mathbf{D},\mathbf{K}} \neq \emptyset$*, we say that* $\mathbf{D}$ *and* $\mathbf{K}$ *are configurable.*

Throughout the remainder of this paper, we assume that $\mathbf{D}$ and $\mathbf{K}$ are configurable. Note that this is always the case when $\mathbf{D}$ and $\mathbf{K}$ are extracted from an empirical data set. More general problems concerning the configurability of arbitrary $\mathbf{D}$ and $\mathbf{K}$ may also be considered. By viewing the annotated bipartite graph $\mathcal{B}(\mathcal{H})$ as a union $\cup_{x\in\mathcal{X}}\mathcal{B}^x(\mathcal{H})$ of single-role bipartite graphs, one for each role x, we can reduce the problem to determining the configurability of the individual pairs $\mathbf{d}^x$ and $\mathbf{k}^x$. An elegant necessary and sufficient condition for the configurability of two integer sequences is provided in early work by Ryser (1960; 2009); Gale (1957).

In $\mathbb{C}_{\mathbf{D},\mathbf{K}}$, each node "remembers" how many times it played role each x and each edge remembers how many nodes playing role x were contained in it. Summing over x, we see that nodes remember their degrees and edges their dimensions. The null space $\mathbb{C}_{\mathbf{D},\mathbf{K}}$ thus generalizes the hypergraph configuration null space of Chodrow (2019a).

A natural approach to defining a null model is to define the uniform measure on $\mathbb{C}(\mathcal{H}_0)$. Such a definition is attractive from a theoretical standpoint, since this measure is also the entropy-maximizing measure on $\mathbb{H}$ subject to the degree and dimension constraints. However, methods for sampling from uniform models suffer from computational issues related to counting edge multiplicities and rejection probabilities, often resulting in slow sampling (Fosdick et al. 2018; Chodrow 2019b). We therefore instead follow the traditional path of Bollobás (1980); Molloy and Reed (1998), and others in defining a configuration model as the output of a stub-matching algorithm.

A *stub* is an indexed pair (v, x, i) of a node and a role; the index i serves simply to distinguish stubs. For each such pair, d_v^x counts the number of times that node v has role x in $\mathcal{H}_0$. We collect all stubs a single set:

$$\Sigma = \bigcup_{x\in\mathcal{X}} \bigcup_{v\in V} \underbrace{\left\{(v, x, 1), \ldots, (v, x, d_v^x)\right\}}_{d_v^x \text{ copies}} \ .$$

Stub-matching forms hyperedges by selecting sets of stubs from the set Σ. For each edge e:

(1) For each role x, uniformly sample k_e^x stubs with role x from Σ, without replacement, and combine them via multiset union. Add the result to the edge set $\mathcal{E}$.

(2) Send $\Sigma \mapsto \Sigma \setminus e$.

The algorithm terminates when it is impossible to form the next edge. When $\mathbf{D}$ and $\mathbf{K}$ are configurable, stub-matching terminates when $\Sigma = \emptyset$ and produces a partition generating a stub-labeled hypergraph with specified degree and dimension sequences.

By construction, the output of stub-matching is distributed according to the uniform measure μ_0 on the set $\Sigma_{\mathbf{D},\mathbf{K}}$ of partitions of indexed stubs into subsets with fixed numbers of stubs per node and elements per subset. Let $g : \Sigma_{\mathbf{D},\mathbf{K}} \rightarrow \mathbb{C}_{\mathbf{D},\mathbf{K}}$ be the map that sends to each such partition its associated hypergraph.

Definition 6 (Configuration Model) *The annotated hypergraph configuration model is the measure* $\mu(\mathcal{H}) = \mu_0(g^{-1}(\mathcal{H})|\mathcal{H}$ *is nondegenerate*).

The configuration model μ weights elements of $\mathbb{C}_{\mathbf{D},\mathbf{K}}$ according to their likelihood of being realized via stub-matching, conditional on nondegeneracy. In this, it differs from the uniform distribution on $\mathbb{C}_{\mathbf{D},\mathbf{K}}$, since the same annotated hypergraph can be realized through multiple configurations. Indeed, any permutation of stubs of the form $(v, x, 1)$, $(v, x, 2)$ does not alter the image under g. Because of this, the configuration model tends to give greater probabilistic weight to annotated hypergraphs in which there are many parallel edges.

By definition, we can in principle sample from μ by repeatedly performing stub-matching until a nondegenerate configuration is obtained, and then applying the function g. This approach is usually impractical, as the probability of realizing a nondegenerate configuration is typically very low. The generalization of limit laws such as those provided by Angel et al. (2016) for a dyadic configuration model governing the probability of nondegeneracy would be a welcome development beyond our present scope.

An alternative approach is to perform stub-matching and then simply discard degenerate edges. This approach was pioneered by Molloy and Reed (1995). The model of Allard et al. (2015) can be used to sample annotated hypergraphs using this method. The theoretical justification of this approach is that, though the probability of at least one degeneracy may be high, the expected *number* of degeneracies is low (see again Angel et al. (2016)). One might therefore suppose that these can be removed without doing significant violence to the topological structure of the realized network. Since our present interest lies in null random graph hypothesis-testing, it is desirable to avoid the issue of nondegeneracy entirely. We do so by developing a simple extension of the standard edge-swap Markov chain, and show that this chain is sufficient to perform approximate sampling from $\mu_{\mathbf{d},\mathbf{k}}$ directly.

Edge-swap Markov chains

Edge-swap Markov Chain Monte Carlo provides an alternative approach to sampling from $\mu_{\mathbf{D},\mathbf{K}}$. The benefit of this method of sampling is that, as long as it is initialized with a non-degenerate hypergraph, all samples produced are guaranteed to be nondegenerate.

It is convenient to define edge swaps on the annotated bipartite graph $\mathcal{B}$. There is considerable literature on edge-swap Markov chains for bipartite graphs (Kannan et al. 1999; Erdős and Gallai 1960; Ryser 1960). The edge-swap Markov chain we develop here may be viewed as a superposition of standard chains, one for each role label. This is in principle sufficient to imply irreducibility and aperiodicity; we provide a full proof in order to make our exposition self-contained.

Definition 7 *An role-preserving edge-swap on $\mathcal{B}$ is a map of pairs of edges:*

$$(v_1, e_1; x), (v_2, e_2; x) \mapsto (v_2, e_1; x), (v_1, e_2; x)$$

We can also regard a role-preserving edge-swap of bipartite edges f_1 and f_2 as a map $\pi : \mathbb{B} \to \mathbb{B}$ that generates a new bipartite graph $\mathcal{B}'$. In this case we write $\mathcal{B}' = \pi(\mathcal{B}|f_1,f_2)$. Note that it is possible that $\mathcal{B} = \pi(\mathcal{B}|f_1,f_2)$; this occurs when $f_1 = (v, e_1; x)$ and $f_2 = (v, e_2; x)$ for some v or $f_1 = (v_1, e; x)$ and $f_2 = (v_2, e; x)$ for some e. Nondegeneracy rules out the case that $f_1 = f_2 = (v, e; x)$ for distinct f_1 and f_2.

Let $\mathcal{B}^x$ denote the edges of $\mathcal{B}$ with role label x. Then, we can construct a Markov chain $\mathcal{B}_t \in \mathbb{B}$ by repeated role-preserving double edge swaps. Some care is needed to ensure that each state of this chain is nondegenerate. The full algorithm is formalized in Algorithm 1. The Markov chain $\mathcal{B}_t$ of hypergraphs induces a chain $\mathcal{H}_t \in \mathbb{H}$ of annotated hypergraphs. Theorem 1 ensures that samples from this chain at sufficiently long intervals will be approximately i.i.d. according to μ.

```
**Algorithm 1** MCMC Sampling for $\mu$
**Input**: Initial annotated hypergraph $\mathcal{H}_0$ with bipartite graph $\mathcal{B}$, sample interval
        $\delta t \in \mathbb{Z}_+$, sample size $s \in \mathbb{Z}_+$.
**Initialization:** $t \leftarrow 0$, $\mathcal{B} \leftarrow \mathcal{B}_0$
**while** $t \leq s(\delta t)$ **do**
    sample $e_1, e_2$ u.a.r. from $\binom{\mathcal{E}_t}{2}$ sample $f_1 = (v_1, e_1; x_1)$ and $f_2 = (v_2, e_2; x_2)$ u.a.r. from $e_1$
    and $e_2$
    **if** $x_1 \neq x_2$ **then**
        pass
    **else**
        $\mathcal{B}' \leftarrow \pi(\mathcal{B}_t|f_1,f_2)$
        **if** *$\mathcal{B}'$ is nondegenerate* **then**
            $\mathcal{B}_{t+1} \leftarrow \mathcal{B}'$ $t \leftarrow t + 1$
        **end**
**end**
**Output**: $\{\mathcal{B}_t$ such that $t|\delta t\}$
```

Theorem 1 *The Markov chain $\mathcal{H}_t \in \mathbb{H}$ is irreducible and reversible with respect to μ.*

Proof Our proof follows the broad contours of those found in Fosdick et al. (2018); Chodrow (2019a). Equivalent and alternative proofs of irreducibility may be found in Kannan et al. (1999); Ryser (2009).

It is convenient to first view stub-matching as an algorithm for generating stub-labeled bipartite graphs. To do so, we observe that the output partition of stub-matching defines a bipartite graph similar to $\mathcal{B}$, except that to each edge $(v, e; x)$ is associated an integer between 1 and d_v^x. Let $\bar{\mathcal{B}}$ denote this stub-labeled bipartite graph, and $\bar{\mathbb{B}}$ the set of such graphs. We recover a standard bipartite graph $\mathcal{B}$ from $\bar{\mathcal{B}}$ by erasing the stub-labels.

We now observe that a bipartite edge-swap $\bar{\mathcal{B}}_{t+1} = \pi(\bar{\mathcal{B}}_t|f_1,f_2)$ always produces a new element of $\mathbb{B}$, since each bipartite edge has a distinct stub-label. For the same reason, if $\bar{\mathcal{B}}_{t+1} = \pi(\bar{\mathcal{B}}_t|f_1,f_2)$, then f_1 and f_2 are the only bipartite edges for which this relation holds. In particular, the transition kernel of the edge-swap Markov chain may be written

$$P(\bar{\mathcal{B}}'|\bar{\mathcal{B}}) = \begin{cases} r(f_1,f_2|\bar{\mathcal{B}}) & \exists f_1,f_2 : \bar{\mathcal{B}}' = \pi(\bar{\mathcal{B}}|f_1,f_2) \\ 0 & \text{otherwise,} \end{cases}$$

where $r(f_1,f_2|\bar{\mathcal{B}})$ is the probability that bipartite edges f_1 and f_2 are sampled and that $x_1 = x_2$. It follows that P will be reversible with respect to the uniform measure on $\mathbb{B}$ provided that $r(f_1,f_2|\bar{\mathcal{B}}) = r(f_1',f_2'|\bar{\mathcal{B}}')$ whenever $\bar{\mathcal{B}}' = \pi(\bar{\mathcal{B}}|f_1,f_2)$ and $\bar{\mathcal{B}} = \pi(\bar{\mathcal{B}}'|f_1',f_2')$. To see why this is the case, note that $r(f_1,f_2|\bar{\mathcal{B}})$ depends on $\bar{\mathcal{B}}$ only through the sizes of edges and the role distributions within each edge. These quantities are preserved under double edge swaps. We thus conclude that $\mathcal{B}_t$ is reversible with respect to μ_0, and therefore $\mathcal{H}_t$ is reversible with respect to μ.

It remains to show irreducibility. We will construct a supported path in state space from $\bar{\mathcal{B}}$ to $\bar{\mathcal{B}}'$, where these are stub-labeled bipartite graphs with fixed marginals. Choose x such that $\bar{\mathcal{B}}^x \setminus \bar{\mathcal{B}}'^x$ is nonempty, and let $(v_1, e_1; x, i_1) \in \bar{\mathcal{B}}'^x \setminus \bar{\mathcal{B}}^x$. Then, there must exist edges of the form $(v_1, e_2; x, i_1)$ and $(v_2, e_1; x, i_2)$ in $\bar{\mathcal{B}}^x \setminus \bar{\mathcal{B}}'^x$, since node v_1 must be connected to some edge with role x, and similarly edge e_1 must be connected to some node with role x in $\bar{\mathcal{B}}$. In particular, the swap $(v_1, e_2; x, i_1), (v_2, e_1; x, i_2) \mapsto (v_1, e_1; x, i_1), (v_2, e_2; x, i_2)$ reduces the size of the set $\bar{\mathcal{B}}^x \setminus \bar{\mathcal{B}}'^x$ by at least one. This is because we have generated the edge $(v_1, e_1; x, i_1)$, which was in $\bar{\mathcal{B}}'^x$ by hypothesis, while neither of the removed edges $(v_1, e_2; x, i_1)$ or $(v_2, e_1; x, i_2)$ were in $\bar{\mathcal{B}}'^x$. Repeating this procedure allows us to reduce the size of $\bar{\mathcal{B}}^x \setminus \bar{\mathcal{B}}'^x$ indefinitely Iterating over all labels x is then sufficient to generate a supported path between any two stub-labeled bipartite graphs $\mathcal{B}, \mathcal{B}' \in \mathbb{B}$. Dropping the stub-labels produces a supported path between any two elements of $\mathcal{B}$, and therefore any two elements of $\mathcal{H}$, as was to be shown. □

Corollary 1 *As $\delta t \to \infty$, the output of Algorithm* 1 *is asymptotically independent and identically distributed according to μ.*

For the feature of guaranteed nondegeneracy, we pay a computational cost in mixing times. While some results are known for mixing times of bipartite edge-swap Markov chains (Kannan et al. 1999; Erdös et al. 2010), the conditions under which these results are obtained are restrictive and often do not hold in empirical data sets. Results available for non-bipartite graphs (Greenhill 2014; 2011) scale poorly with the node degrees and total number of edges, suggesting that the computational burden may be significant. Despite these considerations, edge-swap Markov chains can be nevertheless be practically deployed in a variety of practical settings, see Fosdick et al. (2018) for a review.

Analysis of annotated hypergraphs

In this section, we introduce a series of tools for measuring the structural properties of annotated hypergraphs while flexibly incorporating information about roles. We split these into two categories; those that can be natively defined on the annotated hypergraph structure, and those that can be measured using a projection of the annotated hypergraph to a weighted directed network.

Native polyadic observables

Role densities

The simplest nontrivial statistic is the *individual role density* associated to a node. The individual role density is a probability distribution summarizing the proportion of interactions in which node v plays each role. It may be computed as

$$\mathbf{p}_v = \frac{\mathbf{d}_v}{\langle \mathbf{e}, \mathbf{d}_v \rangle} \,, \tag{4}$$

where $\mathbf{e}$ is the vector of ones, and $\mathbf{d}_v$ is the vector of degrees of node v in each role, defined in Eq. (2).

Local role density

It is of interest to compare the individual role density $\mathbf{p}_v$ of node v to the proportion of roles in a neighborhood of v. Let c_v^y give the number of times in which a node incident to v plays role y, other than v itself. We then have,

$$c_v^y = \sum_{x \in \mathcal{X}} \sum_{e \in \mathcal{E}} \mathbb{I}(\ell(v,e) = x) \sum_{u \in e, u \neq v} \mathbb{I}(\ell(u,e) = y) = \sum_{e : v \in e} k_e^y - d_v^y \,.$$

Normalizing yields the *local role density* $\mathbf{p}'$,

$$\mathbf{p}'_v = \frac{\mathbf{c}_v}{\langle \mathbf{e}, c_v \rangle} \,. \tag{5}$$

The local role density measures not the behavior of node v, but rather the typical behavior of the nodes with which v interacts.

Assortativity

Classically, degree-assortativity measures the tendency of nodes of similar degrees to connect to each other. In particular, it is often observed that nodes of high degree tend to connect to other nodes of high degree. In an annotated hypergraph, each node possesses a distinct degree corresponding to each role. We therefore develop a role-dependent assortativity measure, similar to assortativity for directed graphs. The measure we choose generalizes that of Chodrow (2019a), which was formulated for hypergraphs without annotations. We first provide the mathematical formulation, and then discuss the nature of the correlation it measures.

Fix roles $x, y \in \mathcal{X}$. For any nodes $u, v \in \mathcal{V}$, let

$$s_{uv}^{xy} = \sum_{e \in \mathcal{E}} \mathbb{I}(\ell(u,e) = x)\mathbb{I}(\ell(v,e) = y)$$

count the number of edges in which u has role x and v has role y. We compute an assortativity score as a correlation coefficient between the random variables $d_U^x - s_{UV}^{xy}$ and $d_V^y - s_{UV}^{xy}$, where the random nodes U and V are sampled according to a two-stage scheme. We first select a uniformly random edge e from E, and then select a uniformly random pair of distinct nodes U and V, conditional on the events $\ell(U) = x$ and $\ell(V) = y$. In case e contains no such nodes, we resample it and try again. The resulting probability law for U and V is

$$\mathbb{P}^{xy}(U=u, V=v) = \frac{\sum_{e\in E}\binom{k_e}{2}^{-1}\mathbb{I}(\ell(u,e)=x)\mathbb{I}(\ell(v,e)=y)}{\sum_{e\in E}\binom{k_e}{2}^{-1}\sum_{u',v'\in e}\mathbb{I}(\ell(u',e)=x)\mathbb{I}(\ell(v',e)=y)}\ . \tag{6}$$

To compute a Spearman assortativity coefficient, let q^x_{uv} denote the rank of $d^x_u - s^{xy}_{uv}$ among all pairs u and v. Then, the Spearman assortativity coefficient is given by

$$\rho^{xy} = \frac{\mathrm{cov}\left(q^x_{UV}, q^y_{VU}\right)}{\sqrt{\mathrm{var}\left(q^x_{UV}\right)\mathrm{var}\left(q^y_{VU}\right)}} = \frac{\mathbb{E}\left[\left(q^x_{UV} - \mathbb{E}\left[q^x_{UV}\right]\right)\left(r^y_{VU} - \mathbb{E}\left[q^x_{VU}\right]\right)\right]}{\sqrt{\mathbb{E}\left[\left(q^x_{UV} - \mathbb{E}\left[q^x_{UV}\right]\right)^2\right]\mathbb{E}\left[\left(q^y_{VU} - \mathbb{E}\left[q^y_{VU}\right]\right)^2\right]}}\ , \tag{7}$$

with expectations computed with respect to the law $\mathbb{P}^{xy}$ defined by Eq. (6). By construction, ρ^{xy} is symmetric in the roles x and y. Note also that the definition of q^x_{uv} excludes from the calculation instances in which u and v themselves interact in these roles. The assortativity coefficient supports quantitative investigations of questions like: is it statistically the case that the sender and receiver of a given email tend to send and receive many emails, respectively? Do scholars with many first-authorships tend to collaborate with scholars with many last-authorships?

Since it can in practice be difficult to compute the expectations appearing in Eq. (7) exactly, it is often convenient to estimate them via repeated sampling from Eq. (6). This is the method used in our experiments below.

The weighted projection

We often wish to apply tools from dyadic graph theory and network science to polyadic data. The usual way to do this is to *project* the latter by replacing each k-dimensional hyperedge with a k-clique. When role information is available, we can perform more flexible projections. Let $\mathbf{R} = [r^{xy}] \in \mathbb{R}^{p\times p}$. We refer to $\mathbf{R}$ as a *role-interaction kernel*, which describes directed interaction strengths between pairs of roles. We do not place any restriction on the values in $\mathbf{R}$, although we can without loss of generality rescale to ensure that $\max_{x,y}|r^{xy}| = 1$. In all our examples, the entries of $\mathbf{R}$ will be nonnegative, but in principle negative interaction weights can also be used.

We compute the weighted projection matrix $\mathbf{W} = \mathbf{W}(\mathcal{H}, \mathbf{R})$ entrywise via the formula

$$w_{uv} = \sum_{x,y\in\mathcal{X}} r^{xy} \sum_{e\in\mathcal{E}} \mathbb{I}(\ell(u,e)=x)\mathbb{I}(\ell(v,e)=y)\ . \tag{8}$$

Each entry w_{uv} thus counts the number of edges connecting any pair of nodes, weighted by the entries of $\mathbf{R}$. To illustrate, let us first consider the directed hypergraphs of (Gallo et al. 1993).

Directed hypergraphs possess two possible roles, "source" and "target." To project a directed hypergraph to a weighted dyadic graph that respects the roles, we can compute Eq. (8) using the interaction kernel

$$\mathbf{R} = \begin{array}{c c} & \begin{array}{cc} \text{source} & \text{target} \end{array} \\ \begin{array}{c} \text{source} \\ \text{target} \end{array} & \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array}\right] \end{array}.$$

The result is a weighted directed graph in which w_{uv} counts the number of hyperedges in which u appears as a source and v as a target. This example is somewhat trivial, but the benefit of our general formalism is that flexible modeling choices are possible. For example, in our case study of the Enron email data set below we use the interaction kernel

$$\mathbf{R} = \begin{array}{c} \\ \text{from} \\ \text{to} \\ \text{cc} \end{array} \begin{array}{c} \begin{array}{ccc} \text{from} & \text{to} & \text{cc} \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0.25 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \end{array} .$$

This kernel reflects an assumption that information travels efficiently from senders to direct receivers, but more weakly to cc'd receivers.

Centrality

Standard, dyadic centrality analysis may be performed on the weighted projected graph of an annotated hypergraph. For our examples below, we consider the results of computing Pagerank (Brin and Page 1998) and eigenvector centrality on weighted projected networks (Newman 2010). We note that there are alternative approaches to defining centrality in hypergraphs such as (Zhou et al. 2007) and (Benson 2019). The first of these uses a random walk formulation that can be represented via normalized weighted projections. The latter is applicable only for k-uniform hypergraphs. Our approach, while not a direct generalization of either, is considerably more flexible in applied data analysis.

Modularity and community detection

Many networks display modular or "community" structure, with nodes in the same module displaying higher average rates of connection than nodes in different modules. Of the many extant approaches toward detecting modular structure in networks, modularity maximization (Newman 2006; Newman and Girvan 2004) is one of the most popular. While limitations of the algorithm include a resolution limit (Fortunato and Barthélemy 2006), provable lack of polynomial-time exact algorithms (Brandes et al. 2007), and statistically-restrictive assumptions (Newman 2016), the availability of highly efficient and scalable heuristics (e.g. Blondel et al. (2008)) contribute to its long-standing prevalence in practice.

Let $\mathbf{G} \in \{0,1\}^{n \times \ell}$ denote the one-hot encoding of a partition g. The uth row $\mathbf{g}_i$ of $\mathbf{G}$ is one in entry j iff g assigns node u to group j. The *modularity* of g with respect to a null model η is then

$$Q_\eta(g) = \frac{1}{\langle \mathbf{e}, \mathbf{We} \rangle} \left(\mathbf{G}^T (\mathbf{W} - \mathbb{E}_\eta[\,\tilde{\mathbf{W}}]\,)\mathbf{G} \right) ,$$

where $\mathbf{W}$ is the (weighted) adjacency matrix of the graph under study. The notation $\mathbb{E}_\eta[\,\tilde{\mathbf{W}}]$ refers to the expected realization of a random adjacency matrix $\tilde{\mathbf{W}}$ under the null η. We extend modularity to annotated hypergraphs by using our configuration model as the null and approximating $\mathcal{E}_\mu[\,\tilde{\mathbf{W}}]$ to first order. We note that several other authors have also defined modularity objectives on polyadic data sets. For example, a recent proposal of Kaminski et al. (2018) defines a modularity objective function which directly encodes polyadic relationships. We instead adopt an approach based on dyadic projections. Several papers have taken projection-based approaches to hypergraph partitioning (Kumar et al. 2018; Zhou et al. 2007; Evans and Lambiotte 2009). Our proposed objective may be viewed as an extension of this approach to incorporate role metadata.

Recall that w_{uv} gives the observed weighted edge count from u to v, with the weights specified via $\mathbf{R}$. In order to derive a working notion of dyadic modularity, we need only to estimate the expectation of w_{uv} under a suitably-chosen null. We will approximate this expectation under the annotated hypergraph configuration model.

Let us estimate $\mathcal{E}_\mu\left[M^{xy}_{uv}\right]$, the expected number of edges that contain node u in role x and node v in role y. We begin by forming an edge e via stub-matching. There are k^x_e x-stubs that must be selected to form e. Each of these has probability approximately $\frac{d^x_u}{\sum_\ell d^x_\ell} = \frac{d^x_u}{\langle \mathbf{e}, \mathbf{d}^x \rangle}$ to be node u. Supposing the probability of degeneracy in an individual edge to be small, we can thus approximate the probability that e contains u in role x as $k^x_e \frac{d^x_u}{\langle \mathbf{e}, \mathbf{d}^x \rangle}$. Similarly, the probability that e contains v in role y is $k^y_e \frac{d^y_v}{\langle \mathbf{e}, \mathbf{d}^y \rangle}$. Summing over edges gives our approximation for $\mathcal{E}_\mu\left[M^{xy}_{uv}\right]$:

$$\mathcal{E}_\mu\left[M^{xy}_{uv}\right] = \sum_{e\in E} k^x_e k^y_e \frac{d^x_u d^y_v}{\langle \mathbf{e}, \mathbf{d}^x \rangle \langle \mathbf{e}, \mathbf{d}^y \rangle} \;.$$

We can write this expression more compactly as

$$\mathcal{E}_\mu[\,\mathbf{M}^{xy}] = \frac{\langle \mathbf{k}^x, \mathbf{k}^y \rangle (\mathbf{d}^x \otimes \mathbf{d}^y)}{\langle \mathbf{e}, \mathbf{d}^x \rangle \langle \mathbf{e}, \mathbf{d}^y \rangle} \;,$$

where $\otimes$ denotes the vector outer product.

To compute $\mathcal{E}_\mu[\,\tilde{\mathbf{W}}]$, we weight by $\mathbf{R}$ and sum over role pairs:

$$\mathcal{E}_\mu[\,\tilde{\mathbf{W}}] = \sum_{x,y\in\mathcal{X}} r^{xy} \mathcal{E}_\mu[\,\mathbf{M}^{xy}] \;.$$

We then define a dyadic modularity score of a partition g with one-hot encoding $\mathbf{G}$:

$$Q_\mu(g) = \frac{1}{\langle \mathbf{e}, \mathbf{W}\mathbf{e} \rangle} \mathrm{tr}\left(\mathbf{G}^T(\mathbf{W} - \mathbb{E}_\mu[\,\tilde{\mathbf{W}}]\,)\mathbf{G}\right) \;. \tag{9}$$

As usual, the pre-factor ensures that $-1 \le Q_\mu(g) \le 1$.

It is important to clarify the nature of the null model used in the modularity calculation. A procedure that is commonly followed for studying polyadic data is to construct a projected graph and perform modularity maximization with respect to an implicit null defined over dyadic graphs. In contrast, we have defined a null over the space of annotated hypergraphs, and then computed expectations in the projected graph with respect to this higher-order null. This approach has the benefit of preserving some information about polyadic interactions in the modularity score, even though this score is natively dyadic. These two approaches will generally lead to different null matrices and therefore different partitions.

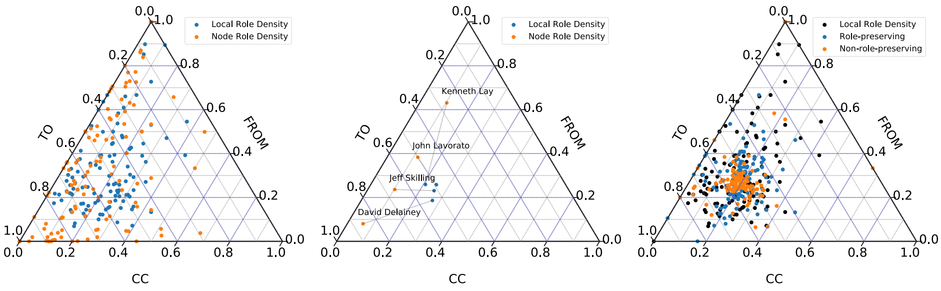
In order to approximately maximize Eq. 9, we adopt the multiway spectral algorithm of Zhang and Newman (2015), though many alternatives are possible. The matrix $\mathbf{W}$ is not symmetric, and therefore it is necessary to make a small adjustment to this algorithm (Leicht and Newman 2008). Rather than computing the leading eigenvectors of the matrix $\mathbf{W} - \mathbb{E}_\mu[\,\tilde{\mathbf{W}}]$, we instead compute the eigenvectors of the symmetrized form $\frac{1}{2}\left(\mathbf{W} + \mathbf{W}^T - \mathbb{E}_\mu\left[\tilde{\mathbf{W}} + \tilde{\mathbf{W}}^T\right]\right)$. The multiway spectral algorithm is then applied to perform this task.

Results

We illustrate how annotated hypergraphs and their null models can be used to enrich analysis of previously studied data sets.

Enron case study

We first focus our analysis on the Enron email data set (Klimt and Yang 2004). This data contains the emails from employees of the company prior to its forced bankruptcy and



(A) (B) (C)

Fig. 1 Role distributions in the Enron data set. **a**. The individual role densities (orange) and local role densities (blue). **b**. The individual role densities (orange) and local role densities (blue) of the four Enron CEOs, now with lines connecting the statistics for each node. **c**. The local role densities (black) compared with a single sample from the role-preserving ensemble (blue) and non-role-preserving ensemble (orange)

shutdown due to corporate fraud and corruption. While this data is temporal in nature we consider only the time-aggregated graph, neglecting the timestamps of edges. We also only consider official email accounts of Enron employees, referrd to as the 'core' group.

Role distributions and assortativity

The data contains three roles: "from", "to", and "cc." These describe the sender, recipients, and carbon copy recipients of emails respectively[1]. Depending on the occupation of the employee, the number of emails they send and receive (and therefore their role participation) will vary. Figure 1a shows the distribution individual and local role densities for the data. Here we see a diverse range of behaviour, with an increased diversity in individual node distributions compared to local distributions. For example, there are nodes that are exclusively message senders or exclusively message receivers, but no node neighbourhoods consist of a single one role. The extent of role hybridization highlights the importance of modeling roles as properties of node-edge pairs, since no node can be assigned a single role.

Figure 1b shows the individual and local role densities for the four CEOs of the company, now each connected by a line. Here we see that, while the CEOs correspond with other individuals who perform similar roles, they exist across a spectrum of behaviour themselves, namely in whether they are senders or receivers of emails. For example, Kenneth Lay was the sender of emails roughly 65% of the time, in comparison with David Delainey, who played that role in less than 10% of interactions. This passive role of David Delainey as a receiver of information rather than originator of it may be reflected in the lesser sentencing he received in comparison to Lay.

In Fig. 1c we show the effect of randomization under hypergraph configuration models. The configuration model preserves the roles of all nodes, while in the non-role-preserving variant we simply erase the role labels. The samples from both null models show a reduction in heterogeneity of the role distribution – randomization has the effect of homogenizing the population. This effect is especially apparent under non-role-preserving randomization and the local role distribution converges towards the average over all nodes.

[1] Blind carbon copy, or "bcc," has been merged into 'cc' for simplicity.

The role assortativity, defined in Eq. (7), captures the tendency of nodes that frequently play one role to be associated with nodes that frequently play another. Figure 2 compares the observed and null-distributed assortativities ρ^{xy} for each combination of roles in the Enron data set. Five combinations are shown – there are six pairs of role labels, and the 'from/from' combination is vacuous since all emails have exactly one sender. In four of the five combinations, the assortativity coefficient is much higher than would be expected under the null and would generally be judged statistically significant. The positive assortativities in the 'from/to' and 'from/cc' combinations quantify the tendency of prolific senders to share information with prolific receivers. The latter combination highlights the importance of using null models to contextualize network measurements. Despite the fact that the observed assortativity is negative, the null distributions are even more so. The data should therefore be judged more assortative than expected by chance. The 'to/to' and 'cc/cc' combinations quantify the tendency of important receivers to do so along the same communication threads. The 'to/cc' combination illustrates the utility of role-preserving randomization – while this measurement would be statistically significant under randomization without role information, it falls within the bulk of the null when roles are preserved. We recall that the Spearman coefficient defined above subtracts out the edges along which u and v interact. It is natural to conjecture that the result for the 'to/cc' combination indicates a tendency for nodes to cluster in recurring 'to/cc' motifs, which are then removed in the course of calculation. An example of a recurring pattern might be frequent emails to an executive with their personal assistant cc'd.

Centralities

In this and subsequent sections, we study the weighted projected graph of the Enron annotated hypergraph. We will primarily use the role interaction kernel given by

$$\mathbf{R} = \begin{array}{c} \\ \text{from} \\ \text{to} \\ \text{cc} \end{array} \begin{array}{c} \begin{array}{ccc} \text{from} & \text{to} & \text{cc} \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0.25 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \end{array} . \tag{10}$$

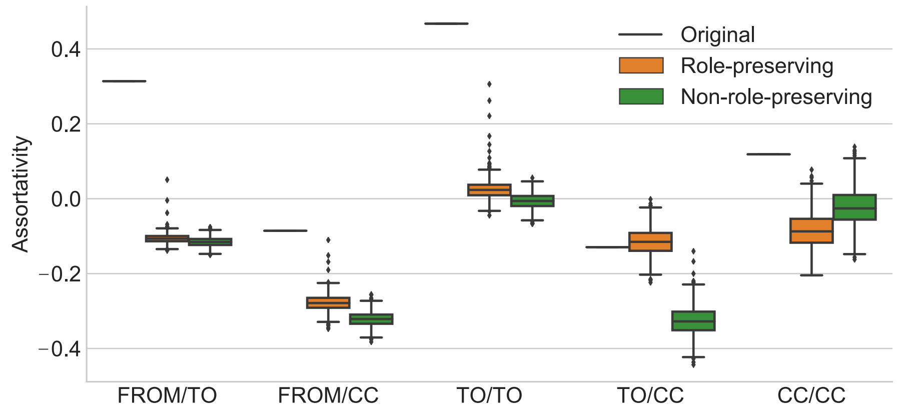


Fig. 2 Hypothesis-testing for significant role-dependent assortativity as measured by Eq. (7). The empirical values (black bars) are shown alongside labeled (orange) and unlabeled (green) configuration null distributions for each combination of roles

This kernel emphasizes flow along edges – information flows strongly to receivers listed in the 'to' field and less strongly to those listed in the 'cc' field.

Figure 3 illustrates the flexibility of interaction kernels in studying graph properties. We compute eigenvector and PageRank centralities on the weighted projected graph using the kernels $\mathbf{R}$ and $\mathbf{R}^T$. High-centrality nodes under $\mathbf{R}$ will tend to be those to whom information flows, while under $\mathbf{R}^T$ they will tend to be those from whom information originates. The kernel $\mathbf{R}$ thus emphasizes information *sinks*, and $\mathbf{R}^T$ information *sources*. In bulk, the source and sink centralities are only weakly correlated, indicating that they capture distinct structural properties of the network. The flexibility of the formalism of annotated hypergraphs with user-specified role interaction kernels supports the discovery of these features.

Modularity maximization

Figure 4 shows the single highest-modularity partition of 100 candidate partitions obtained using this kernel for $k = 4$ communities. Inspection of the clustered adjacency matrix (left) suggests that three of the communities are relatively coherent, while one is highly disperse and essentially serves as a "none of the above" class. On the right, we show the communities themselves, along with nodes colored according to job title.

It is useful to draw contrast against the uniform role interaction kernel, which ignores roles entirely:

$$\mathbf{R}' = \begin{array}{c} \\ \text{from} \\ \text{to} \\ \text{cc} \end{array} \begin{array}{c} \begin{array}{ccc} \text{from} & \text{to} & \text{cc} \end{array} \\ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{array}\right] \end{array} \qquad (11)$$

The use of this kernel produces substantially different communities when compared to the information flow kernel. The best of 100 partitions obtained using this kernel only chose three communities, when up to $k = 4$ were available. This partition achieves modularity $Q = 0.524$. The modularity scores of the two kernel weighting methods should not be compared directly, since they are computed on different modularity matrices.

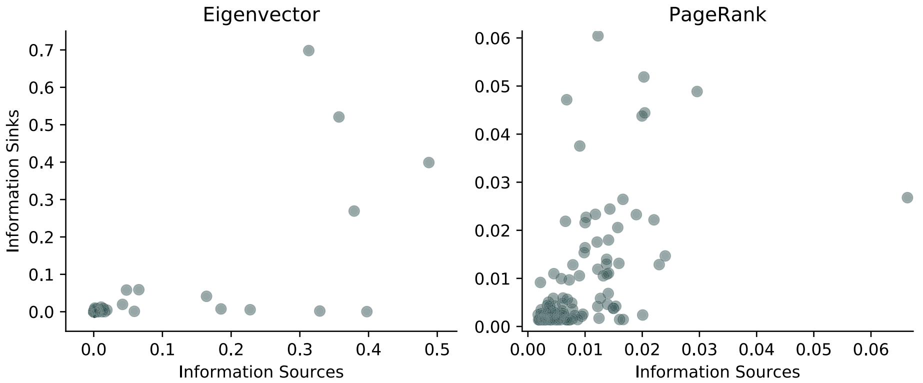


Fig. 3 Sensitivity of eigenvector and PageRank centrality measures to role interaction kernels. The horizontal axis gives centrality scores under the projection $\mathbf{R}^T$, while the vertical gives those under $\mathbf{R}$. The PageRank teleportation parameter is $\alpha = 0.15$

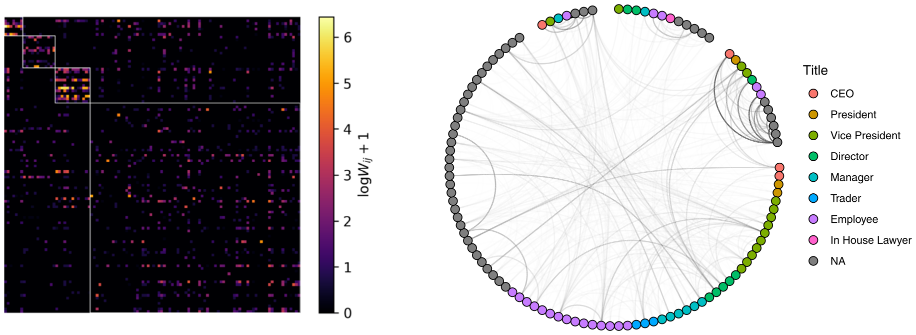


Fig. 4 Example partition of the Enron data set via approximate spectral maximization of Eq. (9). (Left): The adjacency matrix, arranged in order of the partition. Colors give the value of $\log W_{ij} + 1$; log-scaling was chosen for readability. (Right): Visualization of the network. Directed edges have been made undirected. The darkness of each edge corresponds to its weight. The color of each node corresponds to listed job title. In this experiment, $k = 4$, and the kernel used is the information flow kernel R of Eq. (10). The modularity of the shown partition is 0.506

A simple measure of similarity between the two partitions is the normalized mutual information

$$\mathrm{NMI} = 2\frac{I(X,Y)}{H(X)+H(Y)} ,$$

where X and Y are random variables giving the community assignment of a uniformly random node under each scheme, $H(X)$ is the entropy of random variable X, and $I(X, Y)$ the mutual information of the two random variables X and Y. The NMI has maximum value of unity when X and Y are deterministically related, and minimum value of zero when they are statistically independent. In this case, the normalized mutual information between the weighted and unweighted community assignments is 0.55. This score indicates that the partitions are correlated, as we might expect – however, there are nevertheless substantial differences between them. These simple experiments emphasize the importance of appropriately-specified interaction kernels when performing community detection on annotated hypergraphs.

Ensemble study

We now turn to studying annotated graph features systematically across a range of data sets. We consider six data sets, outlined in Table 1, with full descriptions given in Appendix A. We capture the wide variety of possible hypergraphs, those where the number of nodes exceeds that of the number of edges ($|V| \gg |E|$), and those where the number of nodes is much less than the number of edges ($|V| \ll |E|$). The number of node-edge stubs can take a maximum value of $|V||E|$ which corresponds to every node being included in every edge. All the data are social in nature, but diverse in their interpretation. With the exception of the Enron data, all are sparse, with low mean node degree $\langle d \rangle$ and edge dimension $\langle k \rangle$.

For each data set we calculate seven summary statistics. Two of these are the average entropy of the individual and local role densities (Eqs. (4) and (5) resp.), which capture the diversity of roles observed on individual nodes and their neighbourhoods. An entropy of zero indicates no role diversity observed, while maximal entropy[2] indicates all roles are

[2] If the number of roles is $|\mathcal{X}| = p$ then the entropy has a maximum value of $\log_2 p$.

Table 1 Descriptive statistics for each data set

	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{X}\rvert$	$\langle d\rangle$	$\langle k\rangle$
`enron`	112	10,504	3	230.6	2.5
`stack-overflow`	22,131	4,716	3	1.3	6.0
`math-overflow`	410	154	3	1.7	4.6
`scopus-multilayer`	1,677	938	3	1.7	3.1
`movielens`	73,155	43,058	2	2.8	4.7
`twitter`	52,294	123,158	4	5.1	2.2

The fourth and fifth columns give the mean node degree and mean edge dimension, both ignoring role labels

equally likely to be observed. In addition we calculate the mutual information between a node's individual roles and local roles. This quantity is computed as the mean KL-divergence between the individual and local role densities. We also calculate the weighted degree, PageRank, and eigenvector centralities for each node. To create suitable summary statistics we again use entropy, now across the distribution of centrality across nodes. This captured how concentrated the centrality is across all nodes in the hypergraph. Finally we report the number of weakly connected components in the weighted projection.

We assess the significance of each feature by comparing with ensembles generated from both role-preserving and non-role-preserving randomised swaps. We take 500 samples from each null model, each time performing $\lfloor 0.1|\mathcal{E}'|\rfloor$ shuffles, where $|\mathcal{E}'|$ is the number of edge stubs. We use a burn-in period of $10|\mathcal{E}'|$ shuffles to ensure that all chains are sufficiently well-mixed.

In Fig. 5 we show the local role mutual information across each data set. We see in all cases except for `math-overflow` that the local role mutual information is significant when compared to the non-role-preserving null model. This intuitively makes sense given that nodes may switch roles and so any correlation between node states is lost. In all data the local role density is significant when compared to the role-preserving model, however again the `math-overflow` shows the least difference (likely due to the small data size). Despite being significant in all but one data set, the local role mutual information can be both larger than expected (e.g. `enron, scopus-multilayer`) and small than

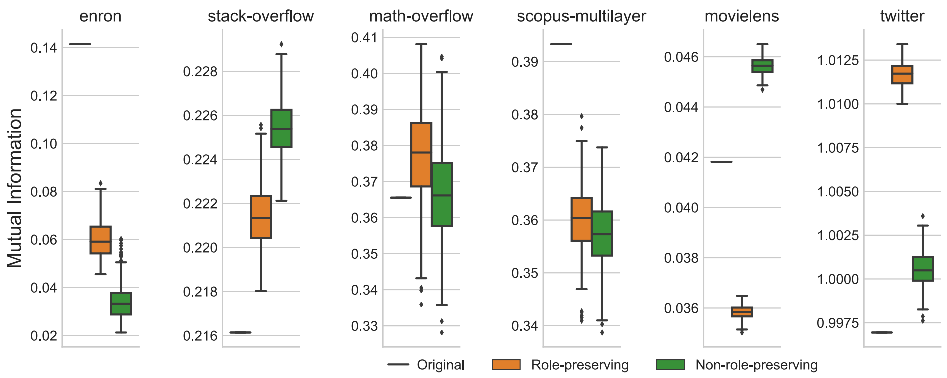


Fig. 5 The average local role mutual information across data sets, with comparisons to the role-preserving (green) and non-role-preserving ensembles (green). In all cases, the observed value falls outside the inter-quartile range for the role-preserving ensemble (orange), and in all but the `math-overflow` data set for the non-role-preserving ensemble (green)

expected (e.g. stack-overflow, twitter) when compared to the null. This can be explained by certain nodes having little diversity of roles in their local neighbourhood. For example, in the enron hypergraph, certain nodes may only ever send messages (and so their neighbourhood is considers of recipients) as we saw in Fig. 1b. However, upon shuffling, these nodes may be swapped with other nodes with more diverse neighbourhoods, effectively increasing the average local diversity. Those which are lower than expected suggest that the hyperedges themselves are relatively uniform in participating roles. In the twitter data this could be due to the limit on the edge cardinality (due to the character limit on posts).

The significance of the remaining features for two data sets is given in Table 2. Naturally the node role entropy is never significant under the role-preserving null model, however in all cases it becomes significant when roles are not preserved upon shuffling. In the enron data set all features (barring the number of connected components) differ significantly from the non-role-preserving null. Interestingly the shuffling effect is not consistent across the different centrality measures. For the eigenvector centrality is more localised than expected (lower entropy) however the PageRank is less localised than expected (higher entropy). Echoing Fig. 1c, the neighbourhood role distributions are more diverse in both null models. For the stack-overflow data, the significance of number of components can easily be explained by the presence of topic cliques in the original data. For example, users answering questions on *Python* may be unlikely to answer questions on *Javascript*. Since the null model is agnostic to these topics, the components are merged under shuffling. The role entropies are all significantly lower than expectation. This can be explained by nodes being consistent in the roles that they take across multiple questions - question answerers tend to answer more questions, for example.

A full summary of the of the analysis across all features and datasets is given in Appendix B.

Table 2 The significance of multiple observables across the enron and stack-overflow data sets

			Non-preserving		Preserving	
Data	Feature	Orig. Val	Avg. Val.	*z*-score	Avg. Val.	*z*-score
enron	Connected components	1.00	1.00	0.00	1.00	0.00
	Local role MI	0.14	0.03	15.68	0.06	9.78
	Local role Entropy	1.24	1.36	-12.47	1.34	-8.24
	Node role Entropy	0.99	1.33	-19.55	0.99	0.00
	Weight. degree entropy	5.87	5.83	3.00	5.87	-0.39
	Weight. eigenvector entropy	4.02	5.84	-154.46	5.88	-55.09
	Weight. pagerank entropy	6.36	6.11	26.47	6.14	13.74
stack-overflow	Connected components	1168.00	1166.71	0.06	1041.01	6.49
	Local role MI	0.22	0.23	-7.69	0.22	-3.88
	Local role Entropy	0.87	0.88	-4.73	0.88	-6.32
	Node role Entropy	0.05	0.07	-19.78	0.05	0.00
	Weight. degree Entropy	13.89	13.81	14.97	13.84	8.71
	Weight. eigenvector entropy	7.95	8.19	-0.38	7.28	1.24
	Weight. pagerank entropy	14.28	14.23	18.13	14.24	13.88

Significance scores coloured red are two standard deviations larger in the original data than the null model. In contrast, those coloured blue are two standard deviations smaller. The results for all data sets are presented in Fig. 6

Discussion

Many of the seminal results in network science were obtained using dyadic, unweighted networks without metadata – perhaps the minimal model of a complex system. As the quantity and richness of networked data have grown, so too has the need to incorporate higher-order interactions and heterogeneous node and edge properties into our models. Toward this program, we have offered annotated hypergraphs as a modeling framework for rich, polyadic data. Annotated hypergraphs may be viewed as a natural extension of directed graphs for modeling heterogeneous, polyadic interactions. Because the setting of annotated hypergraphs is highly general, they allow the data scientist to flexibly incorporate their assumptions about how role information should feature into downstream analysis.

We have also made contributions to the inferential and exploratory analysis of annotated hypergraphs. First, we have formulated a role-aware configuration null model. This model can be used to assess whether an observed structural feature in an annotated hypergraph can be explained purely through information about role-dependent edge and node incidence. Features that cannot may be reasonably attributed to higher-order mechanisms, which may then be investigated. Second, we have provided a small suite of analytical tools that generalize many common methods for studying dyadic networks, including centrality, assortativity, and modularity scores. We have shown how each of these may be used in role-aware ways to highlight diverse features of the data. Along the way, we have argued that the incorporation of metadata enables more detailed approaches to standard methods in network science, while also offering a stable mathematical foundation upon which to build qualitatively new methodologies.

Annotated hypergraphs admit multiple avenues of future work. There are opportunities to define and study diffusion, spreading, and opinion dynamics on annotated hypergraphs, using models that explicitly account for heterogeneous, polyadic interactions. For example, our weighted projection scheme does not incorporate any explicit accounting of edge dimensions. It may be of interest to define a role-dependent simple random walk along the hypergraph. These structures implicitly normalize edge dimensions, implying that a node who received an email along with ten others is in some sense less important than one who received a private communication. This walk could then be used to define alternative measures of centrality and modularity.

Another direction of future development concerns the structure of hyperedges. By assigning roles within each hyperedge, we impose a certain model of how the interaction marked by the edge takes place. Further structural assumptions are possible. In some cases it may be useful to assume, for example, that the hyperedge itself contains a small network between the nodes. The role of the hyperedge in this case is to serve as a single entity housing a network motif (Alon 2007). A null model over such structures would allow for the sampling of random graphs with control over the participation of nodes in various microscale graph structures. While such a model would be substantially more complex, the emerging importance of network motifs (Benson et al. 2016) and network-of-networks modeling (e.g. Kenett et al. (2015)) may provide sufficient impetus to pursue it. The work of Karrer and Newman (2010) offers promising progress in this direction, but we suspect that there is still much to be done. In particular, there is a certain tension between their notion of roles, which is strictly based on the location of a given node in the context of a specified network motif, and ours, which is defined via metadata extrinsic to the network structure. The question of how to reconcile these notions when modeling topological motifs in rich networks is an important one which we hope will receive attention from the community.

Finally, an important feature of many polyadic data sets is that interactions are temporally localized. The incorporation of temporal information into models of hypergraphs and annotated hypergraphs is of substantial importance for modeling realistic dynamics on network substrates. One route may be to generalize temporal event graphs (Mellor 2018) for rich, polyadic data. Such a generalization, along with the development of associated metrics, would be of significant theoretical and practical interest.

This work is a contribution to the project of integrating progressively more complex information into network data science. We foresee that this program will become increasingly important as rich, relational data sets become more readily available.

Appendix A: Data and software

Data sets and choice of role-interaction matrix

Enron email

This consists of the core Enron emails from the archived Enron email database (Klimt and Yang 2004). Core email addresses are those with a valid Enron address. All other emails have been omitted. Here the node roles are *from*, *to*, and *cc* which capture the various fields in a typical email header (in this case bcc has been merged with cc). Note that a node may appear in an edge twice under multiple roles, for example sending a message to oneself.

For this hypergraph we choose the role-interaction matrix to be

$$\mathbf{R} = \begin{array}{c|ccc} & \text{from} & \text{to} & \text{cc} \\ \hline \text{from} & 0 & 1 & 0.25 \\ \text{to} & 0 & 0 & 0 \\ \text{cc} & 0 & 0 & 0 \end{array}.$$

This reflects that information can only be transmitted from the sender. Furthermore this reflects the assumption that information is less likely to be transmitted (or not fully transmitted) to those who are "cc'd."

Scopus multilayer literature

This data consists of academic literature surrounding 'multilayer networks,' collected using Scopus. Specifically we choose all references from the following three reviews and books:

(1) Kivelä et al. "Multilayer networks." Journal of Complex Networks 2.3 (2014): 203-271.

(2) Bianconi. Multilayer networks: structure and function. Oxford University Press, 2018.

(3) Boccaletti et al. "The structure and dynamics of multilayer networks." Physics Reports 544.1 (2014): 1-122.

Authors are assigned roles for each article dependent on their order in the list of authors. Although practices vary across disciplines and institutions, here we distinguish between first, middle, and last authors. When there are fewer than three authors then the role is assigned as first for a single author, and first and last for a pair of authors (regardless of any note of equal contribution).

For this hypergraph we choose the role-interaction matrix to be

$$\mathbf{R} = \begin{array}{c} \\ \text{first} \\ \text{middle} \\ \text{last} \end{array} \begin{array}{c} \begin{array}{ccc} \text{first} & \text{middle} & \text{last} \end{array} \\ \begin{bmatrix} 0 & 1 & 0.5 \\ 0.2 & 0.2 & 0.2 \\ 1 & 0.25 & 0 \end{bmatrix} \end{array}.$$

We make the modelling assumption that the first author is the most knowledgeable and therefore able to spread information to other authors. The last author is often an advisor who can spread information to the first author, while the middle authors can weakly diseminate information between everyone.

MovieLens actor credits

This data contains a list of credits from the MovieLens data collection. We consider the top five billed actors from a collection of [] movies. Actor roles are distinguished by being the top-billed actor, or in the remaining cast.

For this hypergraph we choose the role-interaction matrix to be

$$\mathbf{R} = \begin{array}{c} \\ \text{top} \\ \text{rest} \end{array} \begin{array}{c} \begin{array}{cc} \text{top} & \text{rest} \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0.25 & 0.25 \end{bmatrix} \end{array}.$$

The top billed actor is assumed to be the most diffusive, potentially spreading fame and influence. The lower billed cast have a smaller diffusive rate.

Stack overflow threads

This data contains a list of Stack Overflow question threads which achieved a score greater than 25 between 1st January 2017 to 1st January 2019. The score is calculated and reflects the quality of the question both in terms of its pertinence and its presentation. Here edges reflect questions threads where users can be in three roles. These are the question setter, the question answerers, and the best answerer (chosen by the question setter as the accepted answer). If a question remains unanswered, or an answer has not been accepted, then the answerer and best answerer roles are not present in the edge.

For this hypergraph we choose the role-interaction matrix to be

$$\mathbf{R} = \begin{array}{c} \\ \text{setter} \\ \text{answerer} \\ \text{accepted} \end{array} \begin{array}{c} \begin{array}{ccc} \text{setter} & \text{answerer} & \text{accepted} \end{array} \\ \begin{bmatrix} 0 & 0.1 & 0.1 \\ 0.3 & 0.3 & 0.3 \\ 1 & 0.5 & 0 \end{bmatrix} \end{array}.$$

Here, the question setter may disseminate some information (either about the question, or the topic). The question answerers may share information uniformly across all roles. The node whose question is answered transfers the most information to the question

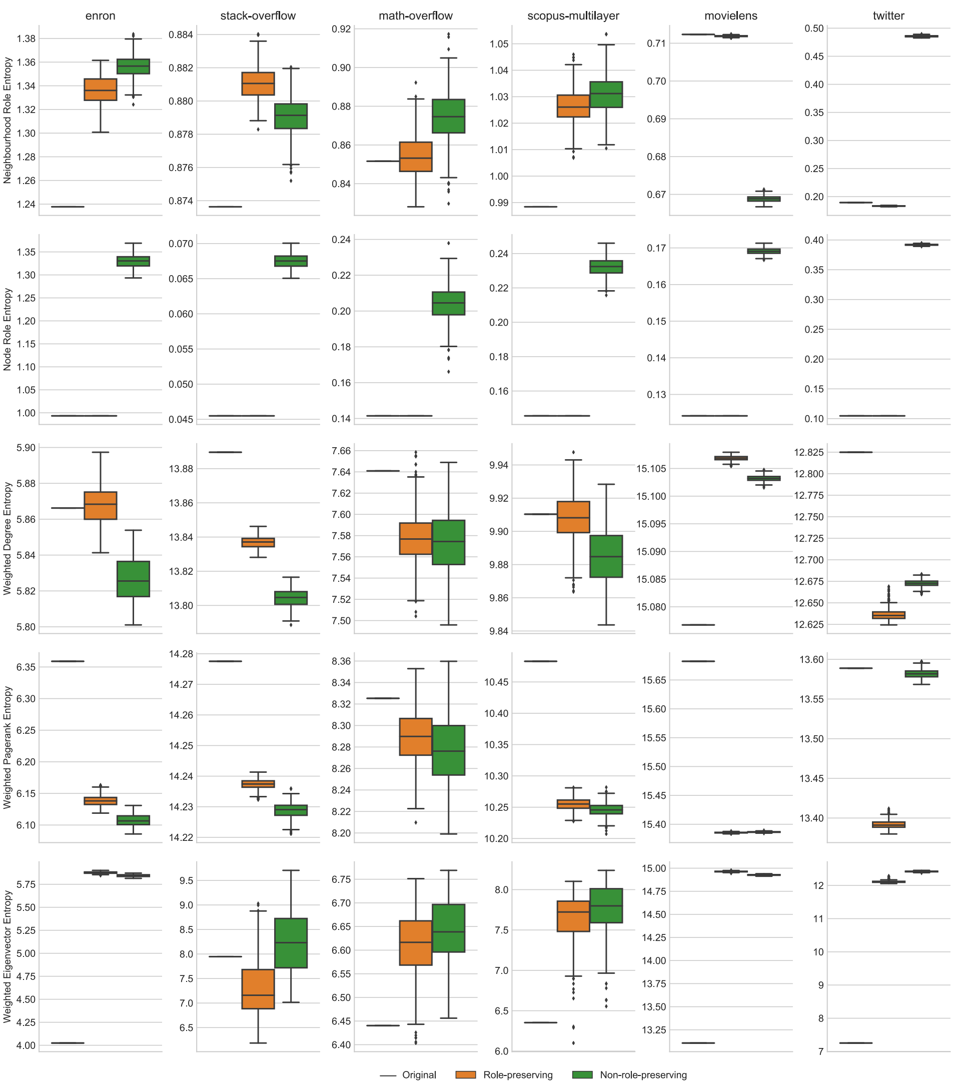


Fig. 6 Full results

setter since the setter has chosen this response to adopt. This accepted answer may also benefit the other answerers with less useful answers.

Math overflow threads

This data is in the identical format to the Stack Overflow threads, however questions threads are now on the topic of mathematical research as opposed to general programming.

Twitter keyword sample

This data contains a list of messages posted on the social media platform Twitter over a 24 hour period. All these messages contained a particular keyword relating to the aviation industry. Each edge corresponds to a message (or *tweet*). Nodes participating in a message can have four possible roles, a sender, a receiver, a retweeter, and the retweeted[3].

[3] See https://help.twitter.com/en/twitter-guide for details of each role.

For this hypergraph we choose the role-interaction matrix to be

$$\mathbf{R} = \begin{array}{c|cccc} & \text{sender} & \text{receiver} & \text{retweeter} & \text{retweeted} \\ \hline \text{sender} & 0 & 1 & 0 & 0 \\ \text{receiver} & 0 & 0 & 0 & 0 \\ \text{retweeter} & 0 & 0.25 & 0 & 0 \\ \text{retweeted} & 0 & 0.25 & 1 & 0 \end{array}.$$

We assume that the sender transmits information to the receivers in a directed fashion (information travelling only one way). A retweeted node can transmit information to the node that retweets it and so to any receivers who are also included in the message.

Appendix B: Ensemble study

In Fig. 6 we present the all results from the ensemble study.

Acknowledgments
The idea for this research was born at the 2019 SIAM Workshop on Network Science, organized by Nina Fefferman and Peter Mucha. PC is grateful to the Mathematical Institute at the University of Oxford for financial support of a collaborative visit.

Authors' contributions
PC and AM contributed equally to the conception of the research, the development of mathematical tools, the design and implementation of computational experiments, and the reporting of results. Both authors read and approved the final manuscript.

Funding
PC is supported by the U.S. National Science Foundation Graduate Research Fellowship under award number 1122374. AM is funded by the Oxford-Emirates Data Science Laboratory.

Availability of data and materials
The datasets generated and analysed during the current study are available from AM's webpage. A Python package for the analysis of annotated hypergraphs is available on Github. This includes the analysis to replicate the figures from this study.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 02139 Cambridge, MA, USA. [2]Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK.

Received: 8 November 2019 Accepted: 9 January 2020
Published online: 30 January 2020

References

Allard A, Hébert-Dufresne L, Young J-G, Dubé LJ (2015) General and exact approach to percolation on random graphs. Phys Rev E 92(6):062807
Alon U (2007) Network motifs: Theory and experimental approaches. Nat Rev Genet 8(6):450–461
Angel O, van der Hofstad R, Holmgren C (2016) Limit laws for self-loops and multiple edges in the configuration model:1–19. arXiv:1603.07172
Battiston F, Nicosia V, Latora V (2014) Structural measures for multiplex networks. Phys Rev E 89(3):032804
Benson AR (2019) Three hypergraph eigenvector centralities. SIAM J Math Data Sci 1(2):293–312
Benson AR, Gleich DF, Leskovec J (2016) Higher-order organization of complex networks. Science 353(6295):163–166
Berge C (1984) Hypergraphs: Combinatorics of Finite Sets, vol. 45. Elsevier
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech-Theory Exp 10:1–12
Bollobás B (1980) A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. Eur J Comb 1(4):311–316
Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, Wagner D (2007) On finding graph clusterings with maximum modularity. In: International Workshop on Graph-Theoretic Concepts in Computer Science. Springer. pp 121–132. https://doi.org/10.1007/978-3-540-74839-7_12
Carlsson G (2009) Topology and data. Bull Am Math Soc 46(2):255–308

Chodrow PS (2019a) Configuration Models of Random Hypergraphs and their Applications. arXiv:1902.09302 [physics, stat]. 1902.09302

Chodrow PS (2019b) Moments of uniform random multigraphs with fixed degree sequences. arXiv preprint arXiv:1909.09037

de Arruda GF, Petri G, Moreno Y (2019) Social contagion models on hypergraphs. arXiv preprint arXiv:1909.11154

Erdös PL, Miklós I, Soukup L (2010) Towards random uniform sampling of bipartite graphs with given degree sequence. arXiv preprint arXiv:1004.2612

Erdős P, Gallai T (1960) Graphs with prescribed degrees of vertices. Mat Lapok 11:264–274

Evans T, Lambiotte R (2009) Line graphs, link partitions, and overlapping communities. Phys Rev E 80(1):016105

Fortunato S, Barthélemy M (2006) Resolution limit in community detection. Proc Natl Acad Sci 104(1):36–41

Fosdick BK, Larremore DB, Nishimura J, Ugander J (2018) Configuring random graph models with fixed degree sequences. SIAM Rev 60(2):315–355

Gale D (1957) A theorem on flows in networks. Pac J. Math 7(2):1073–1082

Gallo G, Longo G, Pallottino S, Nguyen S (1993) Directed hypergraphs and applications. Discret Appl Math 42(2-3):177–201

Gallo G, Scutella MG (1998) Directed hypergraphs as a modelling paradigm. Rivista di matematica per le scienze economiche e sociali 21(1-2):97–123

Ghoshal G, Zlatić V, Caldarelli G, Newman M (2009) Random hypergraphs and their applications. Phys Rev E 79(6):066118

Gomez S, Diaz-Guilera A, Gomez-Gardenes J, Perez-Vicente CJ, Moreno Y, Arenas A (2013) Diffusion dynamics on multiplex networks. Phys Rev Lett 110(2):028701

Greenhill C (2011) A polynomial bound on the mixing time of a markov chain for sampling regular directed graphs. Electron J Comb 18(1):234

Greenhill C (2014) The switch markov chain for sampling irregular graphs. In: Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM. pp 1564–1572. https://doi.org/10.1137/1.9781611973730.103

Greening Jr BR, Pinter-Wollman N, Fefferman NH (2015) Higher-order interactions: understanding the knowledge capacity of social groups using simplicial sets. Curr Zool 61(1):114–127

Heath LS, Sioson AA (2009) Multimodal networks: Structure and operations. IEEE/ACM Trans Comput Biol Bioinforma (TCBB) 6(2):321–332

Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, Koutra D, Faloutsos C, Li L (2012) Rolx: Structural role extraction & mining in large graphs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp 1231–1239

Kaminski B, Poulin V, Pralat P, Szufel P, Theberge F (2018) Clustering via hypergraph modularity:1–17. arXiv:1810.04816

Kannan R, Tetali P, Vempala S (1999) Simple markov-chain algorithms for generating bipartite graphs and tournaments. Random Struct Algoritm 14(4):293–308

Karrer B, Newman ME (2010) Random graphs containing arbitrary distributions of subgraphs. Phys Rev E 82(6):066118

Kenett DY, Perc M, Boccaletti S (2015) Networks of networks–an introduction. Chaos Solitons Fractals 80:1–6

Klamt S, Haus U-U, Theis F (2009) Hypergraphs and cellular networks. PLoS Comput Biol 5(5):1000385

Klimt B, Yang Y (2004) The enron corpus: A new dataset for email classification research. In: European Conference on Machine Learning. Springer. pp 217–226

Kovanen L, Kaski K, Kertész J, Saramäki J (2013) Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc Natl Acad Sci 110(45):18070–18075

Kumar T, Vaidyanathan S, Ananthapadmanabhan H, Parthasarathy S, Ravindran B (2018) Hypergraph clustering: a modularity maximization approach. arXiv:1812.10869

Leicht EA, Newman ME (2008) Community structure in directed networks. Phys Rev Lett 100(11):118703

Marcotte P, Nguyen S (1998) Hyperpath formulations of traffic assignment problems. In: Equilibrium and Advanced Transportation Modelling. Springer. pp 175–200. https://doi.org/10.1007/978-1-4615-5757-9_9

McMorris FR, Warnow TJ, Wimer T (1994) Triangulating vertex-colored graphs. SIAM J Discret Math 7(2):296–306

Mellor A (2018) Event Graphs: Advances and Applications of Second-order Time-unfolded Temporal Network Models. arXiv preprint arXiv:1809.03457

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algoritm 6(2-3):161–180

Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Comb Probab Comput 7(3):295–305

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328(5980):876–878

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582

Newman M (2010) Networks: An Introduction. Oxford University Press, Oxford

Newman MEJ (2016) Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys Rev E 94:052315. https://doi.org/10.1103/PhysRevE.94.052315

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. https://doi.org/10.1103/PhysRevE.69.026113

Brin S, Page L (1998) "The anatomy of a large-scale hypertextual Web search engine"ă(PDF). Comput Netw ISDN Syst 30(1-7):107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. Sci Adv 3(5):1602548

Rotabi R, Danescu-Niculescu-Mizil C, Kleinberg J (2017) Tracing the use of practices through networks of collaboration. In: Eleventh International AAAI Conference on Web and Social Media

Ryser HJ (1960) Matrices of zeros and ones. Bull Am Math Soc 66(6):442–464

Ryser HJ (2009) Combinatorial properties of matrices of zeros and ones. In: Classic Papers in Combinatorics. Springer. pp 269–275. https://doi.org/10.1007/978-0-8176-4842-8_18

Söderberg B (2003) Random graphs with hidden color. Phys Rev E 68(1):015102

Söderberg B (2003) Properties of random graphs with hidden color. Phys Rev E 68(2):026107

Tarnita CE, Antal T, Ohtsuki H, Nowak MA (2009) Evolutionary dynamics in set structured populations. Proc Natl Acad Sci 106(21):8601–8604

Xie J, Qi L (2016) Spectral directed hypergraph theory via tensors. Linear Multilinear Algebra 64(4):780–794

Young JG, Petri G, Vaccarino F, Patania A (2017) Construction of and efficient sampling from the simplicial configuration model. Phys Rev E 96(3):1–6

Zhang X, Newman ME (2015) Multiway spectral community detection in networks. Phys Rev E 92(5):052808

Zhou D, Huang J, Schölkopf B (2007) Learning with hypergraphs: Clustering, classification, and embedding. In: Advances in Neural Information Processing Systems. pp 1601–1608

Publisher's Note

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