

# CONFIGURATION MODELS OF RANDOM HYPERGRAPHS

PHILIP S. CHODROW

**ABSTRACT.** Many empirical networks are intrinsically polyadic, with interactions occurring within groups of agents of arbitrary size. There are, however, few flexible null models that can support statistical inference for such polyadic networks. We define a class of null random hypergraphs that hold constant both the node degree and edge dimension sequences, generalizing the classical dyadic configuration model. We provide a Markov Chain Monte Carlo scheme for sampling from these models, and discuss connections and distinctions between our proposed models and previous approaches. We then illustrate these models through a triplet of applications. We start with two classical network topics – triadic clustering and degree-assortativity. In each, we emphasize the importance of randomizing over hypergraph space rather than projected graph space, showing that this choice can dramatically alter statistical inference and study findings. We then define and study the edge intersection profile of a hypergraph as a measure of higher-order correlation between edges, and derive asymptotic approximations under the stub-labeled null. Our experiments emphasize the ability of explicit, statistically-grounded polyadic modeling to significantly enhance the toolbox of network data science. We close with suggestions for multiple avenues of future work.

Graphs provide parsimonious mathematical descriptions of systems comprised of objects (nodes) and dyadic relationships (edges). When analyzing a given graph, a common task is to compare an observable of interest to its distribution under a suitably specified null model. A standard choice of null for dyadic networks is the class of *configuration models* [5, 9, 33, 16]. Configuration models preserve the degree sequence of the graph, which counts the number of edges incident to each node. These counts are natural first-order statistics of the graph, which are known to constrain many macroscopic graph properties [38]. Preserving these counts gives a natural null model constraint: properties observed in data that are not present in a configuration model require explanation in terms of higher-order graph structures.

In many systems of contemporary interest, groups of arbitrary size may interact simultaneously. Examples include social contact networks [45, 29]; scholarly and professional collaboration networks; [34, 4, 18, 41]; digital communications [27]; classifications on patents [52]; and many more [6]. In the past decades, the dominant approach to these systems has been to represent these networks dyadically, allowing the analyst to apply standard techniques of dyadic network science, including the configuration model. Recent work, however, has highlighted limitations of the dyadic paradigm in modeling of polyadic systems, both in theory [43] and in application domains including neuroscience [20], ecology [23], computational social science [48, 7] among others [6]. The importance of polyadic interactions calls into question the use of the dyadic configuration model in such systems. It is therefore desirable to construct random models for polyadic data that inherit the useful properties of the dyadic configuration model. In this article, we construct two such models on suitably chosen spaces of hypergraphs, and demonstrate their utility for polyadic network data science. Along the way, we argue for two principle theses. First, the choice between dyadic and polyadic null models can determine the directional findings of standard network analyses. Second, the use of polyadic nulls allows the analyst to measure and test rich measures of polyadic structure, thereby expanding the network-scientific toolbox.

---

*Date:* December 17, 2019.

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA,  
pchodrow@mit.edu.

**Outline.** We begin in Section 1 with a survey of the landscape of null models for relational data, including the dyadic configuration model, random hypergraphs, and random simplicial complexes. In Section 2, we define stub- and vertex-labeled configuration models of random hypergraphs. Practical application of these models requires a sampling scheme, which we provide in Section 3. We turn to a triplet of illustrative applications in Section 4. We first consider triadic closure, showing that some networks that would be considered clustered in comparison to dyadic nulls are significantly *less* clustered than the corresponding hypergraph nulls. We then turn to degree-assortativity, where hypergraph data representations allow us to define novel measures and conduct null hypothesis tests. Finally, we introduce a novel measure of correlation between polyadic edges, which can be tested against either the full configuration model or analytic approximations. We close in Section 5 with a summary of our findings and suggestions for future development.

## 1. GRAPHS, HYPERGRAPHS, AND SIMPLICIAL COMPLEXES

Random graph null modeling has a rich history; see [16] for a review. In this section, we take a rapid tour through some of the most important results in configuration-type models for graphs and their generalizations. We begin with a brief review of the configuration model for dyadic graphs.

**Definition 1** (Graph). *A graph  $G = (V, E)$  consists of a finite set  $V$  of nodes or vertices and a multiset  $E$  of pairs of nodes, also called edges. We assume that both sets are endowed with an (arbitrary) order. An edge of the form  $(u, u)$  is called a self-loop. Two distinct edges  $e_1$  and  $e_2$  are parallel if they are equal as sets.*

Let  $n = |V|$  and  $m = |E|$  be fixed. We denote by  $\mathcal{G}^\odot$  the set of all graphs on  $n$  nodes, and by  $\mathcal{G} \subset \mathcal{G}^\odot$  the set of graphs on  $n$  nodes without self-loops. Parallel edges are permitted in  $\mathcal{G}$ . While it is indeed possible to define configuration models on  $\mathcal{G}^\odot$ , we do not do so here [16, 39]. Rather we will formulate most of our results for elements of  $\mathcal{G}$ , only discussing  $\mathcal{G}^\odot$  below in the context of certain technical issues.

The degree sequence of a graph  $G = (V, E)$  is the vector  $\mathbf{d} \in \mathbb{Z}^n$  defined componentwise as

$$(1) \quad d_v = \sum_{e \in E} \mathbb{I}(v \in e) .$$

A configuration model is a probability distribution on the set  $\mathcal{G}_{\mathbf{d}} = \{G \in \mathcal{G} : \deg(G) = \mathbf{d}\}$  of graphs with degree sequence  $\mathbf{d}$ . There are two closely-related model variants which should be distinguished [16]. On its first introduction [9], the configuration model was defined mechanistically through a “stub-matching” algorithm. To perform stub-matching, we place  $d_v$  labeled half-edges (or “stubs”) into an urn for each node  $v$ . We draw half-edges two at a time, with each draw producing an edge. A stub-labeled graph “remembers” which labeled stubs were drawn to form each edge.

**Definition 2** (Stub-Labeled Graphs). *For a fixed node set  $V$  and degree sequence  $\mathbf{d}$ , define the multiset*

$$\Sigma_{\mathbf{d}} = \biguplus_{v \in V} \{v_1, \dots, v_{d_v}\} ,$$

*where  $\uplus$  denotes multiset union. The copies  $v_1, \dots, v_{d_v}$  are called stubs of node  $v$ . A stub-labeled graph  $S = (V, E)$  consists of the node set  $V$  and an edge set  $E = \{\{u_i, v_i\}\}_{i=1}^m$  which partitions  $\Sigma_{\mathbf{d}}$  into unordered pairs. An edge of the form  $\{v_i, v_j\}$  is called a self-loop.*

Let  $\mathcal{S}^\odot$  be the set of stub-labeled graphs, and  $\mathcal{S} \subset \mathcal{S}^\odot$  the set without self-loops. Technically speaking, one should remember that the set  $\mathcal{S}^\odot$  of stub-labeled graphs is not a subset of the set  $\mathcal{G}^\odot$  of graphs, since the objects in the edge-set are of different logical types. The same is true of the sets  $\mathcal{S}$  and  $\mathcal{G}$ . These technical considerations will also apply when we generalize to hypergraphs below, but will not present any major practical issues.

There is a natural surjection  $g : \mathcal{S}^{\mathcal{G}} \rightarrow \mathcal{G}^{\mathcal{G}}$ . If  $S \in \mathcal{S}^{\mathcal{G}}$ ,  $g(S) \in \mathcal{G}^{\mathcal{G}}$  is the graph obtained by replacing each stub  $v_i$  in  $S$  with  $v$  and then consolidating the result as a multiset. We use the notation  $A = g^{-1}(G)$  to refer to the preimage  $A \subseteq \mathcal{S}$  of  $G \subseteq \mathcal{G}$  by  $g$ . We emphasize that  $g$  is not a bijection, and the symbol  $g^{-1}$  should not be interpreted as an inverse of  $g$ . We define  $\mathcal{S}_{\mathbf{d}}^{\mathcal{G}}$  to be  $g^{-1}(\mathcal{G}_{\mathbf{d}}^{\mathcal{G}})$ . Note that an edge  $\tilde{e} \in S$  is a self-loop if and only if  $e \in g(S)$  is. Because of this,  $\mathcal{S} = g^{-1}(\mathcal{G})$ . It is therefore natural to define  $\mathcal{S}_{\mathbf{d}} = g^{-1}(\mathcal{G}_{\mathbf{d}})$ .

**Definition 3** (Dyadic Configuration Models [16]). *Fix  $\mathbf{d} \in \mathbb{Z}_+^n$ . The vertex-labeled configuration model on  $\mathcal{G}_{\mathbf{d}}$  is the uniform distribution  $\eta_{\mathbf{d}}$ . Let  $\lambda_{\mathbf{d}}$  be the uniform distribution on  $\mathcal{S}_{\mathbf{d}}$ . The stub-labeled configuration model on  $\mathcal{G}_{\mathbf{d}}$  is the distribution  $\mu_{\mathbf{d}} = \lambda_{\mathbf{d}} \circ g^{-1}$ .*

In our formalism, the stub-labeled configuration model is not a distribution over the space of stub-labeled graphs  $\mathcal{S}_{\mathbf{d}}$ . Rather, it is the pushforward of such a distribution to the space  $\mathcal{G}_{\mathbf{d}}$  of graphs. Intuitively, the vertex-labeled configuration model assigns the same probability to each graph with degree sequence  $\mathbf{d}$ , while the stub-labeled model weights these graphs according to their likelihood of being realized via stub-matching. One of the key insights of [9], since generalized by works such as [33, 2], is that these two models are related. Let  $\mathcal{G}_{\text{simple}}$  be the set of *simple graphs*, which contain neither self-loops nor parallel edges. Then,  $\mu_{\mathbf{d}}(G | G \in \mathcal{G}_{\text{simple}}) = \eta_{\mathbf{d}}(G | G \in \mathcal{G}_{\text{simple}})$ . Furthermore, when the degree sequence is sampled from a fixed distribution with finite second moment,  $\mu_{\mathbf{d}}(G \in \mathcal{G}_{\text{simple}})$  is bounded away from zero as  $n$  grows large (see, e.g. [2]), implying that repeated sampling from  $\mu_{\mathbf{d}}$  will produce a simple graph in a number of repetitions that is asymptotically constant with respect to  $n$ . As a result, in the “large, sparse regime,” it is possible to sample from the stub-labeled configuration model until a simple graph is obtained, which will then be distributed according to the vertex-labeled model. This relationship is extremely convenient, enabling asymptotic analytic expressions for many quantities of theoretical and practical interest [38].

This close relationship between models is likely the reason why the distinction between them has often been elided in applied network science. Recently, however, the authors of [16] pointed out that, in many data sets, these two models are not interchangeable. It is important to distinguish them when the data may possess multi-edges or self-loops and the edge density is relatively high. The first condition is important because stub- and vertex-labeled models agree only on the subspace of simple graphs, not the full space of multigraphs. The second condition locates us away from the large, sparse regime and implies that parallel edges will occur under stub-matching with non-negligible probability.

From a modeling perspective, the choice of vertex- or stub-labeling must depend on domain-specific reasoning about counterfactual comparisons. Roughly, stub-labeling should be used when, for a fixed graph  $G \in \mathcal{G}$ , the elements of the set  $g^{-1}(G) \subset \mathcal{S}$  have distinct identities in the context of the application domain. This corresponds to asking whether permutations of stubs lead to meaningfully different counterfactual data sets. In contrast, when stub-permutations are either nonsensical or are considered to leave the observed data unchanged, vertex-labeling is to be preferred. For example, in [16], the authors argue that vertex-labeled nulls are most appropriate for studying a collaboration network of computational geometers. Their reason is that stubs in this case correspond to an author’s participation in a paper. It is nonsensical to say that  $A$ ’s first collaboration with  $B$  is  $B$ ’s second collaboration with  $A$ , and therefore stub-labeling is inappropriate.

Configuration models and their variants have played a fundamental role in the development of modern network science. The seminal paper by Molloy and Reed [32] has, according to Google Scholar, been cited at least 2,000 times since its publication, and over 800 times since 2015. How can we extend these models for application to polyadic data sets? A direct approach, taken in early studies such as [34], is to compute the *projected (dyadic) graph*. The projected graph represents each  $k$ -adic interaction as a  $k$ -clique, which contains an edge between each of the possible  $\binom{k}{2}$  pairs of

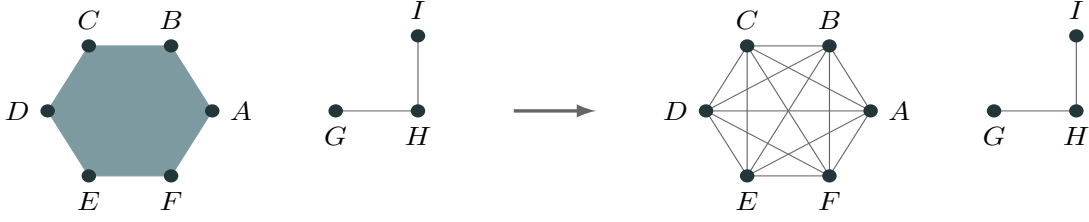


FIGURE 1. A synthetic coauthorship network with  $n = 9$  nodes. On the left, the network represented as a hypergraph with 3 hyperedges. On the right, the projected graph with 17 dyadic edges.

nodes (Figure 1). The resulting dyadic graph may then be randomized according to vertex- or stub-labeled dyadic configuration models. Projecting, however, can have unintended and occasionally counterintuitive consequences. First and most clearly, all properties which depend explicitly on the presence of higher-dimensional interactions are lost. Second, other observables such as node degrees and edge multiplicities may be transformed in undesirable ways; for example, a single interaction between six agents becomes 15 pairwise interactions after projection. As consequence, each of the six agents involved in a single 6-adic interaction are dyadically represented as nodes of degree 5. Third, and most subtly, projecting transforms the null space for downstream hypothesis-testing in ways that may not be intended. For example, projecting the network in Figure 1 prior to randomization implicitly chooses a null space of counterfactuals consisting of 17 two-author papers. This may be undesirable, especially when the null is viewed as a candidate data generating process. Given that the data possesses higher-order interactions, a null model that is by construction unable to produce such interactions may not be physically relevant to the problem at hand.

**Random Hypergraphs.** Considerations such as these motivate the development of dedicated null models for polyadic data. Such models enable the analyst to delay or omit dyadic projection when conducting null-hypothesis testing. We now make a brief survey of efforts to define configuration-type models for polyadic data. Hypergraphs provide the most general context for such models. Hypergraphs are straightforward generalizations of graphs in which each edge is permitted to have an arbitrary number of nodes.

**Definition 4** (Hypergraph). *A hypergraph  $H = (V, E)$  consists of a node set  $V$  and an edge set  $E = \{\Delta_j\}_{j=1}^m$  which is a multiset of multisets of  $V$ . Each subset is called a hyperedge, edge, or, in some contexts, a simplex. Two hyperedges are parallel if they are equal as multisets. A hyperedge is degenerate if it contains two copies of the same node.*

Degenerate hyperedges generalize the notion of self-loops in dyadic graphs. We denote by  $\mathcal{H}^\odot$  the set of all hypergraphs and by  $\mathcal{H}$  the set of all hypergraphs without degenerate edges. As before, parallel edges are permitted in  $\mathcal{H}$ . We continue to let  $n = |V|$  and  $m = |E|$ .

Extant literature provides several approaches to defining null distributions on hypergraphs. One of the earliest approaches [38] takes a somewhat indirect route through bipartite graphs. A bipartite graph contains nodes of two classes, with connections permitted only between nodes of differing classes. To construct a bipartite graph  $B$  from a hypergraph  $H$ , one can construct a layer of nodes in  $B$  corresponding to the nodes  $V$  of  $H$ , and a second layer in  $B$  corresponding to the edges  $E$  of  $H$ . A node  $v$  is linked to an edge-node  $e$  iff  $v \in e$  in  $H$ . We can now apply dyadic configuration models to the randomize  $B$ , before recovering a hypergraph by projecting  $B$  onto its node layer. This approach is natural and convenient, but is only able to gracefully recover a generalization of the stub-labeled configuration model to hypergraphs. Generalizing the vertex-labeled model

requires more complex tools which are not gracefully expressed in the bipartite formalism. We go into greater detail on this connection when discussing sampling methods in Section 3.

A more direct approach is to define a null distribution directly over  $\mathcal{H}$ . In [19], the authors define an analog of the stub-labeled configuration model over the set of hypergraphs in which all edges have three nodes, in the service of studying a tripartite tagging network on an online platform. Somewhat more general models have been formulated for the purposes of community-detection in hypergraphs via modularity maximization, which requires the specification of a suitable null. In [28], the authors develop a degree-preserving randomization via a “corrected adjacency matrix,” which may then be used for modularity maximization on the projected dyadic graph. In [25], the authors explicitly generalize the model of Chung and Lu [12], which preserves degrees in expectation, to non-uniform hypergraphs.

One subspecies of hypergraph has received additional attention. A *simplicial complex* is a hypergraph with additional structure imposed by a subset-inclusion relation: if  $\Delta \in E$ , then  $\Gamma \in E$  for all  $\Gamma \subseteq \Delta$ . Simplicial complexes are attractive tools in studying topological aspects of discrete data [10], since the inclusion condition enables often-dramatic data compression while preserving topological features of interest. Configuration models of simplicial complexes provide one route for conducting null hypothesis tests of such features. The model of [14] achieves analytic tractability by restricting to simplicial complexes with maximal hyperedges of uniform dimension. The authors [53] allow heterogeneous dimensions but sacrifice analytic tractability, instead applying Markov Chain Monte Carlo to sample from the space. In applying any of these models, it is important to remember that subset-inclusion is strong property suited only to certain data-scientific contexts. Particular problems arise when edges possess the semantics of interaction, such as in collaboration networks. Suppose that authors  $A$ ,  $B$ , and  $C$  jointly coauthor a paper. Using hypergraphs, we would represent this collaboration via an edge  $(A, B, C)$ . In the setting of simplicial complexes, on the other hand, subset inclusion would also require us to include the edges  $(A, B)$ ,  $(B, C)$ ,  $(A, C)$ ,  $(A)$ ,  $(B)$ , and  $(C)$ . This may be undesirable, since we are not guaranteed that  $B$  and  $C$ , say, wrote a two-author paper. While simplicial complex modeling may be useful in carefully-selected application areas, in other cases we may require more flexible configuration models defined on more general spaces of polyadic data structures. We now formulate two such models.

## 2. TWO HYPERGRAPH CONFIGURATION MODELS

We now construct two configuration models for general hypergraphs. Our models generalize the stub- and vertex-labeled dyadic configuration models described in the previous section [16].

We use Greek letters to denote random edges of  $H$ , and English letters to denote nonrandom tuples of nodes. For example, the statement  $\Delta = R$  describes the event that a random edge  $\Delta$  has fixed location  $R = (u_1, u_2, u_3, \dots)$ . Let  $\binom{R}{\ell}$  denote the set of all subsets of  $R$  of size  $\ell$ . Let  $\mathbb{I}$  give the indicator function of its argument. We define the *degree sequence*  $\mathbf{d} \in \mathbb{Z}_+^n$  and *dimension sequence*  $\mathbf{k} \in \mathbb{Z}_+^m$  of a hypergraph  $H$  componentwise by

$$d_v = \sum_{e \in E} \mathbb{I}(v \in e) \quad \text{and} \quad k_e = \sum_{v \in V} \mathbb{I}(v \in e).$$

Let  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}^\circ$  and  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}$  denote the sets of hypergraphs with the specified degree and edge dimension sequences with and without degenerate hyperedges, respectively. We say that the sequences  $\mathbf{d}$  and  $\mathbf{k}$  are *configurable* of  $\mathcal{H}_{\mathbf{d}, \mathbf{k}} \neq \emptyset$ .

**Definition 5.** *The vertex-labeled configuration model  $\eta_{\mathbf{d}, \mathbf{k}}$  is the uniform distribution on  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}$ .*

The stub-labeled configuration model is defined similarly as in the dyadic case. The map  $g$  extends naturally to the space of hypergraphs.

**Definition 6.** *Let*

$$\Sigma = \bigsqcup_{v \in N} \{v_1, \dots, v_{d_v}\}$$

*be a multiset of stubs. For each  $v$ ,  $d_v$  copies of  $v$  appear in  $\Sigma$ . A stub-labeled hypergraph  $S$  has as its edge set  $E$  a partition of  $\Sigma$  in which each edge contains at most one stub for each node. Let  $\mathcal{S}_{\mathbf{d}, \mathbf{k}}$  denote the set of all stub-labeled hypergraphs with given degree and dimension sequence, and let  $\lambda_{\mathbf{d}, \mathbf{k}}$  be the uniform distribution on  $\mathcal{S}_{\mathbf{d}, \mathbf{k}}$ . Then, the stub-labeled configuration model is defined by  $\mu_{\mathbf{d}, \mathbf{k}} = \lambda_{\mathbf{d}, \mathbf{k}} \circ g^{-1}$ .*

We have now defined two hypergraph configuration models, generalizing the vertex- and stub-labeled models of [16]. The vertex-labeled configuration model is the entropy-maximizing distribution on  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}$  in the case that the identities of stubs are not meaningful, while the stub-labeled configuration model is the entropy-maximizing distribution when these identities are meaningful. The same considerations discussed in [16] (and briefly in the previous section) apply to the question of when to apply which null model.

### 3. SAMPLING

Stub-matching is a classical method for sampling from the stub-labeled dyadic configuration model [9], and extends naturally to the case of random hypergraphs. Pseudocode for sampling from  $\mu_{\mathbf{d}, \mathbf{k}}$  via stub-matching is provided by Algorithm 1.

---

**Algorithm 1:** Hypergraph Stub-Matching

---

**Input:** Configurable  $\mathbf{d} \in \mathbb{Z}_+^n$  and  $\mathbf{k} \in \mathbb{Z}_+^m$ .  
**Initialization:**  $j \leftarrow 1$ ,  $S \leftarrow \emptyset$ ,  $\Sigma \leftarrow \bigsqcup_{v \in V} \{v_1, \dots, v_{d_v}\}$   
**for**  $j = 1, \dots, m$  **do**  
     $R \leftarrow \binom{W}{k_j}$   
     $W \leftarrow W \setminus R$   
     $S \leftarrow S \cup \{R\}$   
**end**  
**Output:**  $S$

---

Since any stub-labeled graph  $S$  is as likely as any other Algorithm 1, the output, conditioned on nondegeneracy, is distributed according to  $\mu_{\mathbf{d}, \mathbf{k}}$ . As in the dyadic setting, there is nonzero probability for the output of stub-matching to produce a degenerate hypergraph. This probability will generally be large in the presence of highly heterogeneous node degrees – a common phenomenon in empirical data. Many iterations of Algorithm 1 may therefore be necessary in order to generate a single valid sample. Because of this, pure stub-matching is often not a practical method for generating random hypergraphs. That said, the stub-matching algorithm is often useful in proofs involving  $\mu_{\mathbf{d}, \mathbf{k}}$ .

For practical sampling, we consider a Markov Chain Monte Carlo approach, in which we use successive, small alterations to the edge-set  $E$  in order to systematically explore the space  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}$ . Our approach fits within a large class of edge-swap Markov chains [16, 49, 3, 50, 30, 24, 11, 47, 8, 15, 1, 22, 26] used for sampling a wide variety of random structures.

**Definition 7** (Pairwise Reshuffle). *Let  $S \in \mathcal{S}_{\mathbf{d}, \mathbf{k}}$ , and  $\Delta, \Gamma \in S$ . A pairwise reshuffle  $b(\Delta, \Gamma|S)$  of  $\Delta$  and  $\Gamma$  is a sample from the conditional distribution  $\mu(\cdot|E \setminus \{\Delta, \Gamma\})$ . Depending on context, we will regard a pairwise reshuffle as either a random map on stub-labeled hypergraphs or on pairs of hyperedges.*

**Lemma 1.** *Let  $S \in \mathcal{S}$ . Let  $b(\Delta, \Gamma|H) = (\Delta', \Gamma')$  be a pairwise reshuffle which results in  $S' \in \mathcal{S}$ . Then,*

- (1) *The degree and dimension sequences are preserved:  $\deg(S) = \deg(S')$  and  $\dim(S) = \dim(S')$ .*
- (2) *We have  $\Delta' \cap \Gamma' = \Delta \cap \Gamma$ .*
- (3) *Any given realization of  $b$  occurs with probability*

$$(2) \quad q_\mu(\Delta, \Gamma) = 2^{-|\Delta \cap \Gamma|} \binom{|\Delta| + |\Gamma| - 2|\Delta \cap \Gamma|}{|\Delta| - |\Delta \cap \Gamma|}^{-1}.$$

*Proof.* A pairwise reshuffle may be performed via the following sequence, which is an alternative description of the final two iterations (conditioning on nondegeneracy).

- (1) Delete  $\Delta$  and  $\Gamma$  from  $E$ .
- (2) Construct  $\Delta'$  and  $\Gamma'$  as (initially empty) node sets.
- (3) For each node  $v \in \Delta \cap \Gamma$ , add a  $v$ -stub to both  $\Delta'$  and  $\Gamma'$ .
- (4) From the remaining stubs, select  $|\Delta \setminus \Gamma|$  stubs u.a.r. and add them to  $\Delta'$ . Add the remainder to  $\Gamma'$ .
- (5) Add  $\Delta'$  and  $\Gamma'$  to  $E$ .

Each node begins with the same number of edges as it started, so degrees are preserved. Next, by construction,  $|\Delta'| = |\Delta \cap \Gamma| + |\Delta - \Gamma| = |\Delta|$ , and similarly  $|\Gamma'| = |\Gamma|$ . The edge dimension sequence is thus also preserved.

Finally, by construction, step 2 above preserves the intersection  $\Delta \cap \Gamma$ . There are  $2^{|\Delta \cap \Gamma|}$  ways to assign stubs to this intersection. There are a total of  $|\Delta| + |\Gamma| - 2|\Delta \cap \Gamma|$  remaining stubs, and of these one must choose  $|\Delta| - |\Delta \cap \Gamma|$  to be placed in  $\Delta$ . We infer that any given pairwise reshuffle is realized with probability given by Equation (2), as was to be shown.  $\square$

We now define a transition kernel of a first-order Markov chain on the space  $\mathcal{S}_{\mathbf{d}, \mathbf{k}}$ . Write  $S \sim_{\Delta, \Gamma} S'$  if there exists a pairwise shuffle  $b$  such that  $b(\Delta, \Gamma|S) = S'$ . Note that, since each element of each edge has a distinct label in  $\mathcal{S}_{\mathbf{d}, \mathbf{k}}$ , for any  $S$  and  $S'$  there is at most one pair  $(\Delta, \Gamma)$  such that  $S \sim_{\Delta, \Gamma} S'$ . If no such pair exists, we write  $S \not\sim S'$ . Then, let

$$(3) \quad \tilde{p}_\mu(S'|S) = \begin{cases} \binom{m}{2}^{-1} q_\mu(\Delta, \Gamma) & S \sim_{\Delta, \Gamma} S' \\ 0 & S \not\sim S', \end{cases}$$

where  $q_\mu(\Delta, \Gamma)$  is the number of distinct possible shuffles realizable from  $\Delta$  and  $\Gamma$ ; an explicit expression is given in the SI. To sample from  $\tilde{p}_\mu(\cdot|S)$ , it suffices to sample two uniformly random edges from  $E$  and perform a reshuffle. The prefactor  $\binom{m}{2}^{-1}$  gives the probability that any two given edges are chosen.

The sequence  $\{S_t\}$  is Markovian by construction. The following lemma and its corollary ensure that the sequence  $\{H_t\} = \{g(S_t)\}$  is also a Markov chain.

**Lemma 2.** *Let  $H, H' \in \mathcal{H}$ . Suppose that  $S_1, S_2 \in g^{-1}(H)$  and  $S'_1, S'_2 \in g^{-1}(H')$ . Then,  $\tilde{p}_\mu(S'_1|S_1) = \tilde{p}_\mu(S'_2|S_2)$ .*

*Proof.* The objects  $S_1$  and  $S_2$  may each be considered arbitrary stub-labelings of part-edges in  $H$ . Similarly,  $S'_1$  and  $S'_2$  are each arbitrary labelings of part-edges in  $H'$ . However, by Equation (3),  $\tilde{p}_\mu(\cdot|S)$  depends only on the sizes of edges and their intersections in  $H$ , not their labels.  $\square$

**Corollary 1.** *The process  $\{H_t\} = \{g(S_t)\}$  on  $\mathcal{H}_{\mathbf{d}, \mathbf{k}}$  is a Markov chain.*

*Proof.* Markovianity of  $\{H_t\}$  follows from Lemma 2. Indeed, we can construct  $H_t$  mechanistically from  $H_{t-1}$  by choosing  $S_{t-1} \in g^{-1}(H_{t-1})$ , setting  $S_t \sim \tilde{p}_\mu(\cdot|S_{t-1})$ , and then letting  $H_t = g(S_{t-1})$ . Lemma 2 ensures that the distribution of  $H_t$  depends only on  $H_{t-1}$ , and not on the choices of  $S_{t-1}$  and  $S_t$ .  $\square$

**Theorem 1.** *The Markov chain  $\{S_t\}$  on  $\mathcal{S}_{\mathbf{d},\mathbf{k}}$  defined by the kernel  $\tilde{p}_\mu$  is irreducible and reversible with respect to  $\lambda_{\mathbf{d},\mathbf{k}}$ , the uniform distribution on  $\mathcal{S}_{\mathbf{d},\mathbf{k}}$ . If in addition at least two entries of  $\mathbf{k}$  are two or larger,  $\{S_t\}$  is also aperiodic. In this case,  $\lambda_{\mathbf{d},\mathbf{k}}$  is the equilibrium distribution of  $\{S_t\}$ . Furthermore,  $\mu_{\mathbf{d},\mathbf{k}}$  is the equilibrium distribution of the process  $\{H_t\} = \{g(S_t)\}$ .*

*Proof.* We will first show reversibility with respect to  $\lambda_{\mathbf{d},\mathbf{k}}$ . Fix  $S$ . Let  $S \sim_{\Delta,\Gamma} S'$  and  $S' \sim_{\Delta',\Gamma'} S$ . In this case, we have  $\Delta', \Gamma' = b(\Delta, \Gamma|S)$ . Then,

$$\tilde{p}_\mu(S'|S) = \binom{m}{2}^{-1} q_\mu(\Delta, \Gamma) = \binom{m}{2}^{-1} q_\mu(\Delta', \Gamma') = \tilde{p}_\mu(S|S'),$$

as required. The second equality follows from Lemma 1, since  $q(\Delta, \Gamma)$  depends only on  $|\Delta|$ ,  $|\Gamma|$ , and  $|\Delta \cap \Gamma|$ .

Our proof approach for irreducibility generalizes that of [16]. We need to construct a path of nonzero probability between two arbitrary elements  $S_1$  and  $S_2$  of  $\mathcal{S}$ . Let  $E_1$  and  $E_2$  be the edge-sets of  $S_1$  and  $S_2$ , respectively. We first describe a procedure for generating a new stub-labeled hypergraph  $S_3$  such that  $|E_2 \setminus E_3| < |E_2 \setminus E_1|$ . Since  $E_1 \neq E_2$  and  $|E_1| = |E_2|$ , we may pick  $\Delta = \{\delta_1, \dots, \delta_\ell\} \in E_2 \setminus E_1$ . Note that, since  $\Delta \notin E_1$  and the edge dimension sequences must agree, there exists an edge  $\Psi \in E_1 \setminus E_2$  such that  $|\Psi| = |\Delta| = \ell$ . Now, for each  $i$ , since  $\Delta \notin E_1$ ,  $\delta_i$  belongs to a different edge (call it  $\Gamma_i$ ) in  $E_1$ . Note that we may have  $\Gamma_i = \Gamma_{i'}$  in case  $\delta_i$  and  $\delta_{i'}$  belong to the same hyperedge in  $E_1$ . Suppose we have  $j \leq \ell$  such edges. Since  $\delta_i$  is a stub,  $\delta_i$  can belong to only one edge in each hypergraph, and therefore  $\Gamma_k \notin E_2$  for each  $k = 1, \dots, j$ . For each  $k = 1, \dots, j$ , let  $(\Psi_k, \Gamma'_k) = b_k(\Psi_{k-1}, \Gamma_k)$ , where  $b_k$  assigns all elements of the set  $\Delta \cap (\Psi_{k-1} \cup \Gamma_{k-1})$  to  $\Psi_k$  and uniformly distributes the remainder. Since  $\Delta \subseteq (\bigcup_{k=1}^j \Gamma_k)$  by construction, by the end of this procedure we have  $\Psi_j = \Delta$ . Call the resulting stub-labeled hypergraph  $S_3$  with edge set  $E_3$ . Since we have only modified the edges  $\{\Gamma_k\}$  and  $\Psi$ , which are elements of  $E_1 \setminus E_2$ , we have not added any edges to the set  $E_1 \setminus E_2$ , but we have removed one, namely  $\Psi$ . We therefore have  $|E_2 \setminus E_3| < |E_2 \setminus E_1|$ , as desired. Applying this procedure inductively, we obtain a path of nonzero probability between  $S_1$  and  $S_2$ , proving irreducibility.

To prove aperiodicity, we will construct supported cycles of length 2 and 3 in  $\mathcal{S}$ . Since the lengths of these cycles are relatively prime, aperiodicity will follow. To construct a cycle of length 2, pick two edges  $\Delta$  and  $\Gamma$  and any valid reshuffle  $b : (\Delta, \Gamma) \mapsto (\Delta', \Gamma')$ . Then,  $b^{-1} : (\Delta', \Gamma') \mapsto (\Delta, \Gamma)$  is also a valid reshuffle, and the sequence  $(b, b^{-1})$  of transitions constitutes a supported cycle through  $\mathcal{S}$  of length 2. To construct a cycle of length 3, choose two edges  $\Delta$  and  $\Gamma$  which each contain two or more nodes, writing  $\Delta = \{\delta_1, \delta_2, \dots\}$  and  $\Gamma = \{\gamma_1, \gamma_2, \dots\}$ . This is always possible by hypothesis. Then, the following sequence of pairwise reshuffles constitutes a cycle of length 3:

$$\begin{aligned} \{\delta_1, \delta_2, \dots\}, \{\gamma_1, \gamma_2, \dots\} &\mapsto \{\gamma_1, \delta_2, \dots\}, \{\delta_1, \gamma_2, \dots\} \\ &\mapsto \{\gamma_2, \delta_2, \dots\}, \{\delta_1, \gamma_1, \dots\} \\ &\mapsto \{\delta_1, \delta_2, \dots\}, \{\gamma_1, \gamma_2, \dots\}. \end{aligned}$$

We have shown reversibility, irreducibility, and aperiodicity, completing the proof.  $\square$

A small modification enables sampling from the vertex-labeled model  $\eta_{\mathbf{d},\mathbf{k}}$ . Let  $m_\Delta$  give the number of edges parallel to edge  $\Delta$  in hypergraph  $H$ , including  $\Delta$  itself. Define

$$(4) \quad a_\eta(S'|S) = \begin{cases} \frac{2^{|\Delta \cap \Gamma|}}{m_\Delta m_\Gamma} & S \sim_{\Delta,\Gamma} S' \\ 0 & \text{otherwise.} \end{cases}$$

**Theorem 2.** *Let  $\tilde{p}_\eta(S'|S) = a(S'|S)\tilde{p}_\mu(S'|S)$ . Let  $\{S_t\}$  be the Markov chain generated by  $\tilde{p}_\eta$ . Then, the process  $\{H_t\} = \{g(S_t)\}$  is a Markov chain. Furthermore,  $\{H_t\}$  is irreducible and reversible with respect to  $\eta_{\mathbf{d},\mathbf{k}}$ . If in addition  $\mathbf{k}$  has at least two entries larger than 2,  $\{H_t\}$  is aperiodic. In this case,  $\eta_{\mathbf{d},\mathbf{k}}$  is the equilibrium distribution of  $\{H_t\}$ .*



*Proof.* Markovianity of  $\{H_t\}$  follows from the same argument as Corollary 1. Irreducibility and aperiodicity follow from Theorem 1, since the state space  $\mathcal{H}$  is a partition of  $\mathcal{S}$  into equivalence classes induced by  $g$ . It remains to demonstrate reversibility with respect to  $\eta_{\mathbf{d},\mathbf{k}}$ . Let  $p_\eta$  be the transition kernel of  $H_t$ . Fix  $H$  and  $H'$ . Fix  $S \in g^{-1}(H)$  and  $S^* \in g^{-1}(H')$ . Then, we can write

$$\begin{aligned} p_\eta(H'|H) &= \sum_{S' \in g^{-1}(H')} \tilde{p}_\eta(S'|S) \\ &= \sum_{S' \in g^{-1}(H')} a(S'|S) \tilde{p}_\mu(S'|S) \\ &= a(S^*|S) \sum_{S' \in g^{-1}(H')} \tilde{p}_\mu(S'|S). \end{aligned}$$

The expressions appearing in this calculation are independent of the specific choices of  $S$  or  $S^*$  following the same argument as in the proof of Lemma 2. We now evaluate the sum in the third line. The summand is nonzero if and only if  $S \sim_{\Delta,\Gamma} S'$ , in which case its value depends only on  $|\Delta|$ ,  $|\Gamma|$ , and  $|\Delta \cap \Gamma|$ . We therefore count terms. There are  $2^{|\Delta \cap \Gamma|}$  ways to arrange the intersection of  $\Delta$  and  $\Gamma$  in  $\mathcal{S}$ , and  $m_\Delta m_\Gamma$  ways to choose two edges parallel to  $\Delta$  and  $\Gamma$  to reshuffle, all of which generate a distinct element of  $g^{-1}(H')$ . The sum therefore possesses precisely  $a(S^*|S)^{-1}$  terms. We find that  $p_\eta(H'|H) = \tilde{p}_\mu(S'|S)$  for any  $S \in g^{-1}(H)$  and  $S' \in g^{-1}(H')$ . Reversibility of  $p_\eta$  thus follows from reversibility of  $\tilde{p}_\mu$ .  $\square$

Algorithm 2 supplies pseudocode for sampling from the stub- and vertex-labeled hypergraph configuration models.

---

**Algorithm 2:** Markov Chain Monte Carlo for hypergraph configuration models

---

**Input:**  $\mathbf{d}, \mathbf{k}$ , target distribution  $\nu \in \{\mu_{\mathbf{d},\mathbf{k}}, \eta_{\mathbf{d},\mathbf{k}}\}$ , initial hypergraph  $H_0 \in \mathcal{H}_{\mathbf{d},\mathbf{k}}$ , sample interval  $h \in \mathbb{Z}_+$ , desired sample size  $s \in \mathbb{Z}_+$ .

**Initialization:**  $t \leftarrow 0$ ,  $H \leftarrow H_0$

**for**  $t = 1, 2, \dots, sh$  **do**

    sample  $(\Delta, \Gamma)$  u.a.r. from  $\binom{E_t}{2}$

$H' = b(\Delta, \Gamma | H_t)$

**if**  $\text{Uniform}([0, 1]) \leq a_\nu(H'|H)$  **then**

$H_t \leftarrow H'$

**else**

$H_t \leftarrow H_{t-1}$

**end**

**end**

**Output:**  $\{H_t \text{ such that } t|h\}$

---

Theorems 1 and 2 constitute a guarantee that, for sufficiently large sample intervals  $h$ , the hypergraphs sampled from Algorithm 2 will be asymptotically i.i.d. according to the desired distribution. Unfortunately, we are unaware of any mixing-time bounds for this class of Markov chain. It is therefore possible in principle that the scaling in the mixing time as a function of system size is extremely poor, a result suggested by work on related classes of edge-swap Markov chains [21, 22]. Our experience indicates, however, that sampling is possible for configuration models with hundreds of thousands of edges on personal computing equipment in practical time.

**3.1. Connections to Random Bipartite Graphs.** As briefly mentioned in Section 1, a hypergraph  $H = (V, E)$  corresponds in a natural way to a bipartite dyadic graph  $B$ . The graph  $B$  consists of a node set  $V \cup E$ . An edge  $(u, e)$  exists between  $u \in V$  and  $e \in E$  iff  $u \in e$  (in  $H$ ). In this setting, the degree of  $u$  (in  $H$ ) is equal to its degree in  $B$ , and the dimension of  $e$  (in  $H$ ) is similarly equal to its degree in  $B$ . Let  $h$  be the function that assigns to each hypergraph its associated bipartite graph. When both nodes and edges are uniquely labeled,  $h$  is a bijection. It

follows that a probability measure  $\nu$  on the space  $\mathcal{B}_{\mathbf{d},\mathbf{k}}$  of bipartite graphs with node degrees  $\mathbf{d}$  and  $\mathbf{k}$  induces a probability measure  $\nu \circ h^{-1}$  on  $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ . Several extant papers (e.g. [38, 42]) use this equivalence to construct random models of polyadic data. While it is sometimes thought that bipartite randomization supplies a complete solution to null hypergraph sampling, we show in this section that the natural scope of the bipartite method is limited to stub-labeled models.

We first define a bipartite, dyadic, configuration model. We define  $\nu_{\mathbf{d},\mathbf{k}}$  to be the measure on  $\mathcal{B}_{\mathbf{d},\mathbf{k}}$  obtained by performing stub-matching with the node-set  $V \cup E$ , conditioned on the events that (a) all edges have the form  $(u, e)$  for  $u \in V$  and  $e \in E$ , and (b) the bipartite graph is simple, without multi-edges or self-loops. Note that conditioning on the event that  $B$  is simple implies that the stub-labeled and vertex-labeled models are identical in this case. The work of Kannan et al. [26] considers the problem of sampling from  $\nu_{\mathbf{d},\mathbf{k}}$  via bipartite edge-swaps. Such a swap maps  $(u, e), (v, f) \mapsto (u, f), (v, e)$ . By construction, such a swap preserves  $\mathbf{d}$  and  $\mathbf{k}$ . The authors show that a Markov chain which performs successive, random bipartite edge-swaps (while avoiding ones that would lead to a non-simple bipartite graph) is ergodic and therefore sufficient to sample from  $\nu_{\mathbf{d},\mathbf{k}}$ . Such a swap, which viewed in the space  $\mathcal{H}_{\mathbf{d},\mathbf{k}}$  amounts to swapping the edge memberships of nodes  $u$  and  $v$ . Importantly, a sequence of such switches is special case of the pairwise reshuffle Markov chain on  $\mathcal{S}_{\mathbf{d},\mathbf{k}}$ . This implies following relationship:

**Proposition 1.** *The configuration model on simple bipartite graphs is equivalent to the stub-labeled hypergraph configuration model, in the sense that  $\mu_{\mathbf{d},\mathbf{k}} = \nu_{\mathbf{d},\mathbf{k}} \circ h^{-1}$ .*

Proposition 1 makes precise the primary sense in which bipartite randomization provides an approach to random hypergraph modeling. This is a convenient result, since a dyadic edge-swap Markov chain on  $B$  can be used to produce samples from  $\mu_{\mathbf{d},\mathbf{k}}$ . This equivalence may also be used to give alternative proofs of Theorem 1. However, as discussed in [16], many data sets in which we aim to apply null modeling are better represented by vertex-labeled null distributions. There is no obvious route for vertex-labeled sampling through bipartite random graphs. In particular, there is no analogue of Proposition 1 for this case. Thus, even though the work of [26] treats vertex-labeled sampling from  $\mathcal{B}_{\mathbf{d},\mathbf{k}}$ , this does not directly suffice for vertex-labeled sampling from  $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ . The reason is that sampling from the vertex-labeled measure  $\eta_{\mathbf{d},\mathbf{k}}$  requires adjusting for permutations of parallel hyperedges. When  $H$  contains multiple hyperedges of dimension three or greater, it is necessary to track multiple node-edge incidence relations in order to check when hyperedges are parallel.

It is possible to write down a version of Algorithm 2 for vertex-labeled sampling in which the fundamental data structure is a bipartite graph rather than a hypergraph. However, the result would not, to the author’s knowledge, correspond to any standard random bipartite graph model. Expressing both models directly on the space  $\mathcal{H}_{\mathbf{d},\mathbf{k}}$  of hypergraphs supports both conceptual clarity and a convenient formulation of MCMC for both stub- and vertex-labeled models. Incidentally, we note that this discussion constitutes another, separate setting in which adherence to dyadic methods can limit our data-analytical horizons. An exclusive focus on nulls realizable through bipartite methods obscures the possibility of vertex-labeled polyadic models.

#### 4. NETWORK ANALYSIS WITH RANDOM HYPERGRAPHS

We now illustrate the application of hypergraph configuration models through three simple network analyses. We first study triadic closure in polyadic networks, finding that the use of polyadic nulls can generate directionally different, interpretable study findings when compared to dyadic nulls. We then turn to degree-assortativity, defining and testing three distinct measures of association via polyadic data representations and randomizations. Finally, we study the tendency of edges to intersect on multiple vertices in the **email-Enron** data set, finding using simulation and analytical methods that large intersections occur at much higher rates than would be expected by random chance. Collectively, these cases illustrate the use of polyadic methods to define and

analyze richer measures of network structure, and the use of polyadic nulls in interpreting the results.

The data sets for case study were gathered, cleaned, and generously made public by the authors of [6]. In certain experiments, data were temporally filtered in order to reduce their size; these cases of have been explicitly noted in the text and the filtering procedure described in Appendix A. Importantly, in no case was the filtering operation motivated by the expense of Monte Carlo sampling; rather, the bottlenecks were standard, expensive computations such as triangle-counting in dyadic graphs.

**4.1. Triadic Closure.** Triadic closure refers to the phenomenon that, in many networks, if two nodes  $u$  and  $v$  interact with a third node  $w$ , then it is statistically likely that  $u$  and  $v$  also interact with each other. Studies such as [46, 34, 38] observed triadic closure in many empirical networks, and highlighted the fact that dyadic configuration models tend to be unable to reproduce this behavior. Traditionally, triadic closure is measured by a ratio of the number of triangles (closed cycles on three nodes) that are present in the graph, compared to the number of “wedges” (subgraphs on three nodes in which two edges are present).<sup>1</sup> Local and global variants of this ratio have been proposed. We follow the choice of [46] and work with the *average local clustering coefficient*. Let  $T_v$  denote the number of triangles incident on  $v$ , and  $W_v$  the number of wedges. Note that  $W_v = \binom{d_v}{2}$ . The average local clustering coefficient is

$$(5) \quad \bar{C} = \frac{1}{|N|} \sum_{v \in N} \frac{T_v}{W_v}.$$

It is direct to show [37] that, in dyadic configuration models and under mild sparsity assumptions,  $\bar{C}$  decays to zero as  $n$  grows large.

The average local clustering coefficient  $\bar{C}$  is a natively dyadic metric, in the sense that “wedges” and “triangles” are defined explicitly in terms of 2-edges. To compute  $\bar{C}$  in polyadic data, it is therefore necessary to project a hypergraph down to a dyadic graph. In the context of hypothesis-testing, there is some subtlety involved in the choice of when to do this. One method is to project first and then randomize via a dyadic null model. This is the most common historical approach, used for example in [38]. Alternatively, one may randomize via polyadic nulls prior to projection. This approach has the effect of preserving clustering induced by polyadic edges, since an edge of dimension  $k$  contains  $3\binom{k}{3}$  wedges and  $3\binom{k}{3}$  ordered triangles.

Table 1 summarizes a sequence of experiments performed on two collaboration networks (top) and four communication networks (bottom). For each network, we computed the observed local clustering coefficient  $\bar{C}$  on the unweighted projected graph. We then compared the observed value to its null distribution under four randomizations. We first randomized using the vertex- and stub-labeled hypergraph configuration models, *prior* to projecting and measuring  $\bar{C}$ . These results are shown in the second and third columns. We then reversed the order, first computing the projected graph and randomizing via dyadic configuration models. The results are shown in the fourth and fifth columns.

Benchmarking against dyadic configuration models yields mixed results. Vertex-labeled configuration models conclude in all cases that the observed degree of clustering is significantly higher than would be expected by random chance. Stub-labeled benchmarking concludes that **congress-bills** and the two email data sets have significantly less clustering than expected, while the remainder have significantly more. The stub-labeled results should be approached with caution – for reasons discussed in detail in [16], the stub-labeled configuration model is a less-relevant comparison for these data sets than the vertex-labeled model.

Hypergraph randomization leads to directionally different conclusions. First, the expected values of  $\bar{C}$  under both hypergraph vertex- and stub-labeled nulls are much closer together than under

<sup>1</sup>Recent measures have been developed for higher-order notions of clustering on larger subgraphs; see [51].

	$\bar{C}$	Hypergraph		Projected	
		Vertex	Stub	Vertex	Stub
<code>congress-bills*</code>	0.608	0.601(1)	0.622(2)	0.451(2)	0.611(1)
<code>coauth-MAG-Geology*</code>	0.8200	0.8196(7)	0.8186(7)	0.00035(3)	0.00035(3)
<code>email-Enron</code>	0.658	0.825(3)	0.808(4)	0.638(5)	0.797(3)
<code>email-Eu*</code>	0.540	0.569(4)	0.601(4)	0.398(4)	0.598(4)
<code>tags-ask-ubuntu*</code>	0.571	0.609(4)	0.631(5)	0.183(4)	0.499(6)
<code>threads-math-sx*</code>	0.293	0.435(3)	0.426(3)	0.041(1)	0.093(2)

TABLE 1. Average local clustering coefficients for selected data sets, compared to their expectations computed under vertex- and stub-labeling of hypergraph and projected graph models. Parentheses show standard deviations in the least-significant figure under the equilibrium distribution of each null model. Starred\* data sets have been temporally filtered as described in Appendix A.

dyadic nulls, indicating that the polyadic statistical test is much less sensitive than the dyadic test to the choice of vertex- and stub-labeling. Second, the vertex-labeled null separates the two collaboration networks from the four communication networks. These two data sets are only slightly more clustered than expectation under the vertex-labeled model, and only `congress-bills` would be considered “significantly more clustered” than its expectation under most  $p$ -value based tests. The stub-labeled model also has expectation close to the observed values, but finds `congress-bills` to be slightly-but-significantly less clustered than would be expected by chance, while the significance of `coauth-MAG-Geology` would depend strongly on the desired power of the test. In contrast, the four communication networks are all significantly *less* clustered than either vertex- or stub-labeled nulls would expect. Not only is there no clustering beyond that implied by the edge dimensions; triadic closure even appears to be inhibited in these data sets.

From a purely statistical perspective, these examples highlight the importance of careful null model selection in hypothesis-testing for triadic closure. More physically, use of hypergraph nulls allows us in this case to distinguish data sets by their generative mechanisms. The communication networks are all less clustered than expected, while the collaboration networks are approximately as clustered as expected. This result is to some extent intuitive. Collaborations between many agents often have nontrivial coordination costs that scale with the number of agents involved. It may be easier to assemble and coordinate a set of overlapping groups than a single large collective. In such cases, one may expect to observe clustering near or above that expected at random, since overlaps between related groups would generate triangles. In contrast, in digital communications it is essentially effort-free to construct interactions between larger groups of agents. Examples include adding an email address to the “cc” field or introducing participants to thread on a forum. In such cases, triangles composed of distinct edges are energetically unnecessary, and may reflect redundant information flow. We therefore hypothesize that these systems have a tendency to absorb potential triangles into higher-dimensional interactions. This results in lower levels of clustering than would be expected under polyadic nulls. These considerations hint toward the importance of studying edge correlations via natively polyadic metrics as we do in Section 4.3.

**4.2. Degree-Assortativity.** A network is degree-assortative when nodes of similar degrees preferentially interact with each other. Early studies found that different categories of social, biological, and technological networks display different patterns of assortative mixing by degree [36, 35, 13].

Social networks, for example, are frequently measured to be degree-assortative. In this context, degree-assortativity is often taken to indicate a tendency for popular or productive agents to interact with each other.

We measure degree-assortativity in hypergraphs via a generalization of the standard Spearman rank assortativity coefficient to hypergraphs. Importantly, there are multiple possible generalizations, each of which measures distinct structural information about degree correlations. Let  $E_{\geq 2} = \{\Delta \in E : |\Delta| \geq 2\}$ . Let  $h : E_{\geq 2} \rightarrow N^2$  be a (possibly random) *choice function* that assigns to each edge  $\Delta$  two distinct nodes  $u, v \in \Delta$ . Three possibilities of interest are:

$$\begin{aligned} \text{(Uniform)} \quad & h(\Delta) = (u, v) \sim \text{Uniform}\left(\binom{\Delta}{2}\right) \\ \text{(Top-2)} \quad & h(\Delta) = (u, v) = \underset{w, w' \in \binom{\Delta}{2}}{\operatorname{argmax}} d_w d_{w'} \\ \text{(Top-Bottom)} \quad & h(\Delta) = (u, v) = \left( \underset{w \in \Delta}{\operatorname{argmax}} d_w, \underset{w \in \Delta}{\operatorname{argmin}} d_w \right) \end{aligned}$$

The Uniform choice function selects two distinct nodes at random. The Top-2 choice function selects the two distinct nodes in the edge with largest degree. The Top-Bottom choice function selects the nodes with largest and smallest degree.

Let  $r : N \rightarrow \mathbb{R}$  be a ranking function on the node set; we will always take  $r(u)$  to be the rank of node  $u$  by degree in the hypergraph. For fixed  $h$ , let  $f : E \rightarrow \mathbb{R}^2$  be defined componentwise by  $f_j(\Delta) = (r \circ h_j)(\Delta)$ . Then, the *generalized Spearman assortativity coefficient* is the empirical correlation coefficient between  $f_1(\Delta)$  and  $f_2(\Delta)$ :

$$(6) \quad \rho_h = \frac{\sigma^2(f_1(\Delta), f_2(\Delta))}{\sqrt{\sigma^2(f_1(\Delta), f_1(\Delta)) \sigma^2(f_2(\Delta), f_2(\Delta))}},$$

where  $\sigma^2(X, Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$  and brackets express averages over pairs of edges in  $E$ .

In the case of dyadic graphs, the three choice functions above are trivially identical, since there is only one way to pick two nodes from an edge of size two. On polyadic data, however, the resulting Spearman coefficients capture usefully different classes of information. For example, in studying coauthorship networks, they may be used to test hypotheses such as the following:

- (1) **Generic Assortativity:** On a given paper, most coauthors will simultaneously be more or less prolific than average.
- (2) **Junior-Senior Assortativity:** The least prolific author on a paper will tend to be relatively more prolific if the most prolific author is relatively more prolific.
- (3) **Senior-Senior Assortativity:** The two most prolific authors on a paper will tend to be simultaneously more or less prolific than average.

While the corresponding Spearman coefficients may in general be correlated, substantial variation manifests across study data sets. Figure 2 shows measurements and significance tests for one synthetic data set and the six empirical data sets studied in the previous section. The synthetic data consists of five copies of the hypergraph shown in Figure 1. For each data set, we compute the dyadic assortativity coefficient on the projected graph (first row), as well as each of the three polyadic assortativity coefficients defined above.

The synthetic data (first column) illustrates a stark case in which dyadic hypothesis-testing leads to a finding of statistically-significant assortativity, while polyadic hypothesis-testing finds statistically-significant *disassortativity*. In each of the empirical data sets, the dyadic and polyadic tests show directional agreement. However, the polyadic tests highlight several features of the data missed by the dyadic tests. In the two email data sets, all four coefficients are positive and to the right of the null distributions, though projecting (first row) increases the significance of the coefficients relative to hypergraph randomization (second row). The two forum data sets (`threads-math-sx` and `tags-ask-ubuntu`) are disassortative when compared to vertex-labeled

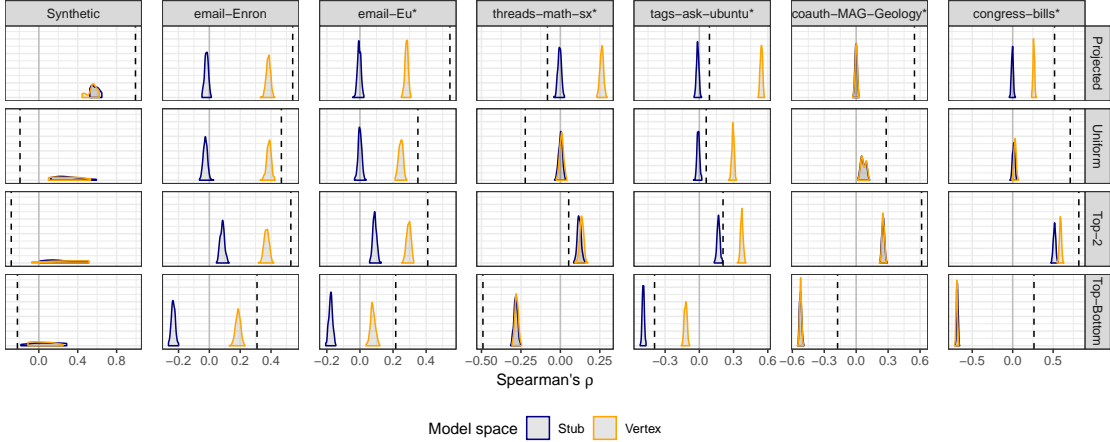


FIGURE 2. Significance tests of degree-assortativity in synthetic and empirical networks. The synthetic data consists of five copies of the illustrative network shown in Figure 1. In each figure, the dashed line gives the observed Spearman correlation, and densities give the null distributions under vertex- and stub-labeled configuration models. In the first row only, the hypergraph was projected down to an unweighted dyadic graph prior to randomization. Starred\* data sets have been temporally filtered as described in the SI.

nulls. The fact that `tags-ask-ubuntu` is disassortative despite a positive uniform hypergraph Spearman coefficient speaks to the importance of carefully specified null hypothesis testing. Interestingly, the misspecified stub-labeled randomization would lead to the opposite finding. The coauthorship network `coauth-MAG-Geology` is highly assortative in all metrics – including the top-bottom measure, which is negative. The `congress-bills` data set is also assortative in all measures. Unlike the other data sets, the uniform hypergraph coefficient lies farther from the bulk of its null distribution than does the projected coefficient. We note in passing that, whereas the stub-labeled and vertex-labeled hypergraph distributions had similar expected rates of triadic closure Table 1, their distributions of degree-assortativity coefficients vary substantially, and would in some cases lead to directionally different study conclusions.

When studying triadic closure, we saw how hypergraph null models could lead us to differently-contextualize standard graph metrics. When studying assortativity, we gain even more. Use of hypergraph nulls allows us to forgo the dyadic projection operation, and thereby define rich polyadic assortativity measures. Hypergraph null models thus enable us to expand our network-analytic toolboxes by measuring and interpreting novel structural patterns in polyadic data.

**4.3. Hyperedge Intersection Profiles.** Let  $\Delta, \Gamma \in H$ . What is the size of their intersection? In the case of dyadic graphs, the intersection can have size at most two, when  $\Delta$  is parallel to  $\Gamma$ . In hypergraphs intersections of arbitrary sizes may occur. The existence of large intersections in a data set may indicate the emergence of polyadic social ties between groups of agents, or interpretable event sequences such as email threads or series of related scholarly papers. Several recent papers [6, 40] have studied similar questions by considering the rate at which “holes” in the hypergraph tend to be “filled in” by higher-order interactions. We take a simpler approach, defining a measure which is both easily computed and amenable to analytical approximation.

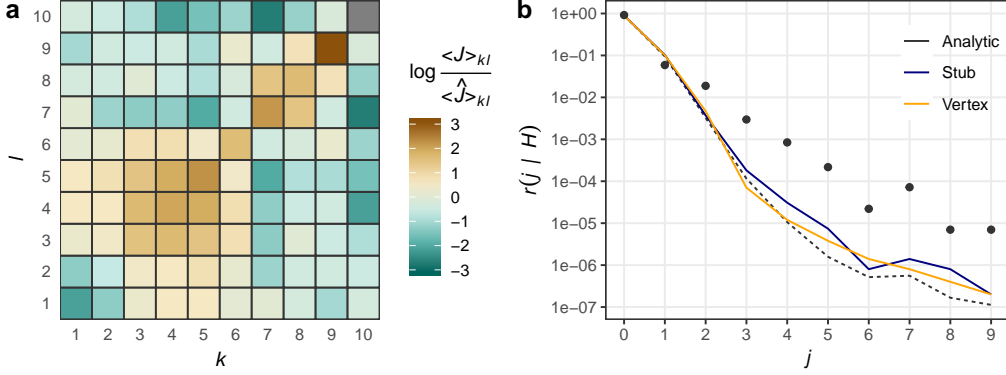


FIGURE 3. Analysis of intersection profiles in the **email-Enron** data set. (a): The average of the intersection size normalized by the expectation  $\langle \hat{J} \rangle_{k\ell}$  under the vertex-labeled configuration model. Positive values indicate that the data displays larger intersections than expected under the configuration model for the corresponding values of  $k$  and  $\ell$ . Colors are shown on a log scale. The missing value at  $(k, \ell) = (10, 10)$  indicates that no nonempty intersections were observed between edges of these sizes in the Monte Carlo sampling runs. (b): Marginal intersection profile (points) of the empirical data, compared to null distributions under the stub- and vertex-labeled configuration models. The dashed gray line gives the analytic approximation of Equation (7). Note the logarithmic vertical axis.

**Definition 8** (Intersection Profile). *For fixed  $k, \ell \in \mathbb{Z}_+$ , the conditional intersection profile of a hypergraph  $H \in \mathcal{H}$  is the distribution*

$$r_{k\ell}(j|H) = \langle \mathbb{I}(|\Delta \cap \Gamma| = j) \rangle_{k\ell},$$

where  $\langle \cdot \rangle_{k\ell}$  denotes the empirical average over all hyperedges  $\Delta$  of size  $k$  and  $\Gamma$  of size  $\ell$ . The marginal intersection profile is

$$r(j|H) = \langle \mathbb{I}(|\Delta \cap \Gamma| = j) \rangle,$$

with the average taken over all pairs of distinct edges in  $E$ .

Large values of  $r_{k\ell}(j|H)$  indicates that edges of size  $k$  and  $\ell$  frequently have intersections of size  $j$  in  $H$ . Empirical data sets may possess complex patterns of correlation between edges of various sizes. Evaluating whether an observed conditional or marginal intersection profile is noteworthy requires comparison to appropriately-chosen null models.

Figure 3 demonstrates the use of hypergraph configuration models to study the intersection profile of the **email-Enron** data set. In Figure 3(a), we compare the empirical average intersection size  $\langle J \rangle_{k\ell} = \sum_{j=0}^{\infty} j r_{k\ell}(j|H)$  to its average  $\langle \hat{J} \rangle_{k\ell}$  under the vertex-labeled configuration model. Higher values of the ratio  $\frac{\langle J \rangle_{k\ell}}{\langle \hat{J} \rangle_{k\ell}}$  indicate the presence of denser intersections between edges of sizes  $k$  and  $\ell$ . Notably, the empirical averages are not uniformly higher than the null model averages, even on the diagonal. There is apparent block structure, indicating that edges of certain sizes tend to correlate most strongly with certain other sizes. Edges of dimension 3 through 6 tend to interact strongly with each other, as do edges of dimension 7 and 8. However, edges in the smaller group interact more weakly with edges in the larger group than would be expected by chance. Further, more detailed study may be able to shed light on the groups of agents involved in these overlapping communications.

Figure 3(b) gives a global view of the data using the marginal intersection profile. The observed profile (points in Figure 3(b)) is nearly linear on semilog axes through  $j = 6$ , suggesting that the decay in the intersection size is roughly exponential. In order to evaluate whether this behavior indicates nonrandom clustering between edges, we again turn to hypergraph configuration models. The expectation  $\hat{r}(j) = \mathbb{E}_\nu[r(j|H)]$  of the marginal intersection profile under a configuration model  $\nu \in \{\mu_{\mathbf{d},\mathbf{k}}, \eta_{\mathbf{d},\mathbf{k}}\}$  measures the typical behavior of a comparable random hypergraph. The solid lines in Figure 3(b) give these expected profiles under both stub- and vertex-labeled models, which directionally agree. The observed data shows fewer intersections on single vertices than would be expected by chance. On the other hand, for  $j \geq 3$ ,  $r(j|H)$  exceeds  $\hat{r}(j)$  by an order of magnitude or more, suggesting substantial higher-order correlation in the data. These results likely reflect the passing of multiple messages between the same sets of users.

Some data sets may be too large to practically estimate  $\hat{r}(j)$  by Monte Carlo methods. In such cases, it is possible to approximate  $\hat{r}(j)$  under the stub-labeled configuration model analytically, using the following asymptotic result.

**Theorem 3.** Fix  $\ell$ ,  $k$ , and  $j$ . Let  $\mathbf{D} \in \mathbb{Z}_+^n$  be a vector of i.i.d. copies of positive, discrete random variable  $D \in \mathbb{Z}_+$  such that  $D \leq d_{\max}$  almost surely for some  $d_{\max}$ . Let  $\mathbf{K} \in \mathbb{Z}_+^n$  be any vector of edge dimensions configurable with  $\mathbf{D}$ . Let  $H \sim \mu_{\mathbf{D},\mathbf{K}}$ , and let  $\Delta$  and  $\Gamma$  be uniformly random edges of  $H$ . Then, with probability approaching unity as  $n$  grows large,

$$(7) \quad \hat{r}_{k\ell}(j) = (1 + O(n^{-1})) j! \binom{k}{j} \binom{\ell}{j} \left( \frac{1}{n} \frac{\mathbb{E}[D^2] - \mathbb{E}[D]}{\mathbb{E}[D]^2} \right)^j.$$

*Proof.* Let  $\langle d \rangle = \frac{1}{n} \sum_{u \in N} d_u$  denote the empirical mean degree of a given degree sequence  $\mathbf{d}$ . Assume without loss of generality that  $\Delta = \{\delta_1, \dots, \delta_k\}$  and  $\Gamma = \{\gamma_1, \dots, \gamma_\ell\}$  are the first two hyperedges formed by Algorithm 1, conditioned on nondegeneracy. There are  $\binom{k}{j}$  ways to choose the  $j$  elements of  $\Delta$  contained in  $\Delta \cap \Gamma$ , and similarly  $\binom{\ell}{j}$  ways to choose the elements of  $\Gamma$ . There are then  $j!$  ways to place these two sets in bijective correspondence. Define the event  $A = \{\delta_h = \gamma_h, h = 1, \dots, j\}$ . Then,  $\hat{r}_{k\ell}(j) = j! \binom{k}{j} \binom{\ell}{j} \mu_{\mathbf{D},\mathbf{K}}(A)$ . To compute  $\mu_{\mathbf{D},\mathbf{K}}(A)$ , we may explicitly enumerate

$$\mu_{\mathbf{D},\mathbf{K}}(A) = \sum_{u \in N} \frac{d_u}{n\langle d \rangle} \frac{d_u - 1}{n\langle d \rangle - 1} \left[ \sum_{v \in N \setminus \{u\}} \frac{d_v}{n\langle d \rangle - d_u} \frac{d_v - 1}{n\langle d \rangle - d_u - 1} \left[ \sum_{w \in N \setminus \{u,v\}} \dots \right] \right],$$

with a total of  $j$  sums appearing. In each summation, the first factor gives the probability that  $\delta_1 = u$  and the second that  $\gamma_1 = u$ . Consider the innermost summation, which may be written

$$(8) \quad S_R = \sum_{z \in N \setminus R} \frac{d_z}{n\langle d \rangle - \sum_{y \in R} d_y} \frac{d_z - 1}{n\langle d \rangle - \sum_{y \in R} d_y - 1}$$

for a set  $R$  of size  $j - 1$ . Since  $D \leq d_{\max}$  a.s., we may employ Chebyshev's inequality to find that  $(n\langle d \rangle)^{-1} \sum_{y \in R} d_y = O(n^{-1})$  w.h.p. We may therefore w.h.p. expand both factors within Equation (8), obtaining the expression

$$\sum_{z \in N \setminus R} \frac{d_z(d_z - 1)}{n^2 \langle d \rangle^2} (1 + O(n^{-1})).$$

Using Chebyshev's inequality again, we also obtain asymptotic behavior on the other expressions appearing above.  $\langle d \rangle = (1 + O(n^{-1})) \mathbb{E}[D]$  and  $\sum_{z \in N \setminus R} \frac{d_z(d_z - 1)}{n} = (1 + O(n^{-1})) (\mathbb{E}[D^2] - \mathbb{E}[D])$ , both w.h.p. We have therefore shown that

$$(9) \quad S_R = (1 + O(n^{-1})) \left( \frac{1}{n} \frac{\mathbb{E}[D^2] - \mathbb{E}[D]}{\mathbb{E}[D]^2} \right)$$

w.h.p. This argument may be repeated inductively for each of the remaining  $j - 1$  sums, each of which contributes the same factor appearing in Equation (9), proving the theorem.  $\square$



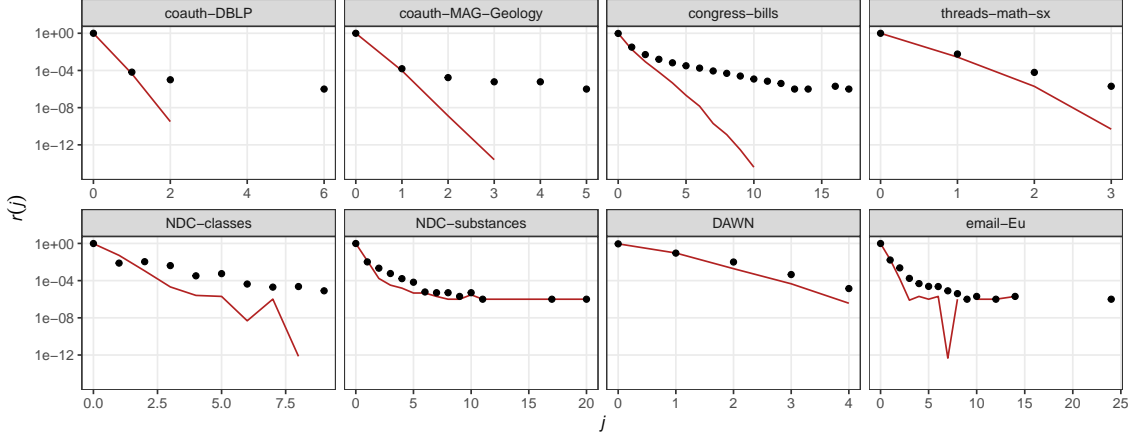


FIGURE 4. Points give the observed intersection profile for six large polyadic data sets. The solid line gives the null intersection profile of Theorem 3. In this visualization, full data sets were used – the temporal filtering described in Appendix A was not performed.

Figure 3 shows the resulting approximation for  $\hat{r}_{k\ell}$  as a dashed line, finding excellent qualitative agreement. This approximation may be used to study intersection profiles in data sets of arbitrary size. Figure 4 shows the use of this approximation to study intersection profiles in hypergraph data sets of arbitrary size. The top set of panels shows four data sets in which the approximate null intersection profile consistently underestimates the rates of large intersections by several orders of magnitude, clearly indicating the presence of correlation structure over and above what would be expected under hypergraph randomization. The lower panels show four additional data sets in which the approximate null profile more closely approximates the observed data.

## 5. DISCUSSION

Configuration models of random hypergraphs preserve the first moments of the data – the degree and edge-dimension sequences – while remaining maximally ignorant about additional data structure. These models extend the widely-used configuration models for dyadic graphs, and serve as natural null models for polyadic network data analysis. We have demonstrated how to define, sample from, analyze, and apply these models. We have seen that the choice between nulls can greatly impact the directional findings of studies of empirical polyadic data. The analyst faced with such a choice must therefore carefully consider whether dyadic simplification will lead to data representations and null spaces that are relevant for their application area. Second, employing polyadic nulls often allows the analyst to define novel measures that can illuminate higher-order structure in data. We have illustrated this with extended assortativity measures and intersection profiles, but many more extensions are possible. We hope that the introduction of statistically-grounded hypergraph nulls will encourage analysts to design, measure, and carefully interpret many novel measures of polyadic network structure.

There are several directions of future work on configuration models of random hypergraphs. Beginning with theory, many classical asymptotic results on dyadic configuration models invite generalization. These include probabilistic characterization of component sizes; cycles and parallel edges; and the diameter of the connected component in various regimes. We also highlight two applications of potential interest. The first is motif analysis. A network motif is a subgraph that appears with higher-than-expected frequency in a given network [31], relative to a given null

model. Considering the explicit dependence of this definition on the null, we conjecture that motif-discovery algorithms based on polyadic nulls may highlight importantly distinct structure when compared to dyadic nulls. A second promising application is in hypergraph clustering and community detection. A recent paper [25] offers a definition of modularity — a common quality function for network partitioning — based on a polyadic generalization of the Chung-Lu model [12]. In this case, the modularity of a given partition may be computed analytically. The same calculations used to prove Theorem 3 can also be used to show that the stub-labeled configuration model will give an asymptotically equivalent expression. However, for the large class of data sets more appropriately modeled by vertex-labeled nulls, other methods may be necessary. We anticipate that pursuing these tasks will pose interesting theoretical and computational challenges.

**Funding.** This work was supported by the National Science Foundation Graduate Research Fellowship under award number 1122374.

**Acknowledgments.** I am grateful to Patrick Jaillet for helpful discussions from which this work benefited substantially.

**Software.** A `hypergraph` class, written in Python 3.5, is available at

<https://github.com/PhilChodrow/hypergraph>.

This class includes implementations of Monte Carlo sampling for both stub- and vertex-labeled configuration models.

**Data.** The data sets used in this paper were prepared by the authors of [6] and accessed from <https://www.cs.cornell.edu/~arb/data/>.

#### APPENDIX A. DATA PREPARATION

The data sets used in this paper were prepared by the authors of [6] and accessed from <https://www.cs.cornell.edu/~arb/data/>. Some data sets have been filtered to exclude edges prior to a temporal threshold  $\tau$  in order to promote practical compute times on triangle counting and mixing of vertex-labeled models in projected graph spaces. Notably, in no cases was sampling from hypergraph configuration models the computational bottleneck. Thresholds were chosen to construct data with edge sets of approximate size  $m \approx 10^4$ , but are otherwise arbitrary. Temporal data subsets were used in the generation of Table 1 and Fig. 2. Table 2 gives the node and edge counts of both the original data and the data after temporal subsetting when applicable.

	Original		$\tau$	Filtered	
	$n$	$m$		$n$	$m$
email-Enron	143	10,886	—	—	—
email-Eu	1,006	235,264	$1.105 \times 10^9$	817	32,117
congress-bills [17, 18]	1,719	260,852	$7.315 \times 10^5$	537	6,661
coauth-MAG-Geology [44]	1,261,130	1,591,167	2017	73,436	23,434
threads-math-sx	201,864	719,793	$2.19 \times 10^{12}$	11,880	22,786
tags-ask-ubuntu	200,975	192,948	$2.6 \times 10^{12}$	2,120	19,338

TABLE 2. Summary of data preparation. When  $\tau$  is given, the filtered data set consists in all edges that occurred after time  $\tau$ .

## REFERENCES

- [1] AMANATIDIS, G., GREEN, B. & MIHAIL, M. (2015) Graphic realizations of joint-degree matrices. *arXiv preprint arXiv:1509.07076*, pp. 1–18.
- [2] ANGEL, O., VAN DER HOFSTAD, R. & HOLMGREN, C. (2016) Limit laws for self-loops and multiple edges in the configuration model. *arXiv:1603.07172*, pp. 1–19.
- [3] ARTZY-RANDRUP, Y. & STONE, L. (2005) Generating uniformly distributed random networks. *Phys. Rev. E*, **72**, 056708.
- [4] BARABASI, A.-L., JEONG, H., NEDA, Z., RAVASZ, E., SCHUBERT, A. & VICSEK, T. (2002) Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, **311**(3-4), 590–614.
- [5] BENDER, E. A. & CANFIELD, E. R. (1978) The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, **24**(3), 296–307.
- [6] BENSON, A. R., ABEBE, R., SCHAUB, M. T., JADBABAIE, A. & KLEINBERG, J. (2018) Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, **115**(48), 11221–11230.
- [7] BENSON, A. R., GLEICH, D. F. & LESKOVEC, J. (2016) Higher-order organization of complex networks. *Science*, **353**(6295), 163–166.
- [8] BLITZSTEIN, J. & DIACONIS, P. (2011) A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet mathematics*, **6**(4), 489–522.
- [9] BOLLOBÁS, B. (1980) A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, **1**(4), 311–316.
- [10] CARLSSON, G. (2009) Topology and data. *Bulletin of the American Mathematical Society*, **46**(2), 255–308.
- [11] CARSTENS, C. J. (2015) Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Physical Review E*, **91**(4), 042812.
- [12] CHUNG, F. & LU, L. (2002) The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, **99**(25), 15879–15882.
- [13] COLIZZA, V., FLAMMINI, A., SERRANO, M. A. & VESPIGNANI, A. (2006) Detecting rich-club ordering in complex networks. *Nature Physics*, **2**(2), 110–115.
- [14] COURTNEY, O. T. & BIANCONI, G. (2016) Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Physical Review E*, **93**(6), 1–26.
- [15] DEL GENIO, C. I., KIM, H., TOROCZKAI, Z. & BASSLER, K. E. (2010) Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PloS one*, **5**(4), e10012.
- [16] FOSDICK, B. K., LARREMORE, D. B., NISHIMURA, J. & UGANDER, J. (2018) Configuring random graph models with fixed degree sequences. *SIAM Review*, **60**(2), 315–355.
- [17] FOWLER, J. H. (2006a) Connecting the Congress: A Study of Cosponsorship Networks. *Political Analysis*, **14**(04), 456–487.
- [18] FOWLER, J. H. (2006b) Legislative cosponsorship networks in the U.S. House and Senate. *Social Networks*, **28**, 454–465.
- [19] GHOSHAL, G., ZLATIĆ, V., CALDARELLI, G. & NEWMAN, M. E. J. (2009) Random hypergraphs and their applications. *Physical Review E*.
- [20] GIUSTI, C., GHRIST, R. & BASSETT, D. S. (2016) Two’s company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. *Journal of Computational Neuroscience*, **41**(1), 1–14.
- [21] GREENHILL, C. (2011) A polynomial bound on the mixing time of a Markov chain for sampling regular directed graphs. *The Electronic Journal of Combinatorics*, **18**(1), 234.
- [22] ——— (2014) The switch Markov chain for sampling irregular graphs. in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1564–1572. SIAM.
- [23] GRILLI, J., BARABÁS, G., MICHALSKA-SMITH, M. J. & ALLESINA, S. (2017) Higher-order interactions stabilize dynamics in competitive network models. *Nature*, pp. 210–213.
- [24] JERRUM, M. & SINCLAIR, A. (1990) Fast uniform generation of regular graphs. *Theoretical Computer Science*, **73**(1), 91–100.
- [25] KAMINSKI, B., POULIN, V., PRALAT, P., SZUFEL, P. & THEBERGE, F. (2018) Clustering via hypergraph modularity. *arXiv:1810.04816*, pp. 1–17.
- [26] KANNAN, R., TETALI, P. & VEMPALA, S. (1999) Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures & Algorithms*, **14**(4), 293–308.
- [27] KLIMT, B. & YANG, Y. (2004) Introducing the Enron Corpus. in *CEAS*.
- [28] KUMAR, T., VAIDYANATHAN, S., ANANTHAPADMANABHAN, H., PARTHASARATHY, S. & RAVINDRAN, B. (2018) Hypergraph clustering: a modularity maximization approach. *arXiv:1812.10869*.
- [29] MASTRANDREA, R., FOURNET, J. & BARRAT, A. (2015) Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, **10**(9), 1–26.
- [30] MCKAY, B. D. & WORMALD, N. C. (1990) Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, **11**(1), 52–67.

- [31] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. & ALON, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**(5594), 824–827.
- [32] MOLLOY, M. & REED, B. (1995) A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, **6**(2-3), 161–180.
- [33] ——— (1998) The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability, and Computing*, **7**(3), 295–305.
- [34] NEWMAN, M. E. J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, **64**(1), 1–8.
- [35] ——— (2002) Assortative mixing in networks. *Physical Review Letters*, **89**(20), 1–5.
- [36] ——— (2003) Mixing patterns in networks. *Physical Review E*, **67**(2), 13.
- [37] ——— (2010) *Networks: An Introduction*. Oxford University Press.
- [38] NEWMAN, M. E. J., STROGATZ, S. H. & WATTS, D. J. (2001) Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, **64**(2), 17.
- [39] NISHIMURA, J. (2018) The connectivity of graphs of graphs with self-loops and a given degree sequence. *Journal of Complex Networks*, **6**(6), 927–947.
- [40] PATANIA, A., PETRI, G. & VACCARINO, F. (2017) The shape of collaborations. *EPJ Data Science*, **6**(1), 1–16.
- [41] PORTER, M. A., MUCHA, P. J., NEWMAN, M. E. J. & WARMBRAND, C. M. (2005) A network analysis of committees in the U. S. House of Representatives. *Proceedings of the National Academy of Sciences*, **102**(20), 7057–7062.
- [42] SARACCO, F., DI CLEMENTE, R., GABRIELLI, A. & SQUARTINI, T. (2015) Randomizing bipartite networks: the case of the World Trade Web. *Scientific Reports*, **5**(10595), 1–18.
- [43] SCHAUB, M. T., BENSON, A. R., HORN, P., LIPPNER, G. & JADBABAIE, A. (2018) Random walks on simplicial complexes and the normalized Hodge Laplacian. *arXiv:1807.05044*, pp. 1–36.
- [44] SINHA, A., SHEN, Z., SONG, Y., MA, H., EIDE, D., HSU, B.-J. P. & WANG, K. (2015) An Overview of Microsoft Academic Service (MAS) and Applications. in *Proceedings of the 24th International Conference on World Wide Web*. ACM Press.
- [45] STEHLÉ, J., VOIRIN, N., BARRAT, A., CATTUTO, C., ISELLA, L., PINTON, J. F., QUAGGIOTTO, M., VAN DEN BROECK, W., RÉGIS, C., LINA, B. & VANHEMS, P. (2011) High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, **6**(8), 1–13.
- [46] STROGATZ, S. H. & WATTS, D. J. (1998) Collective dynamics of “small-world” networks. *Nature*, **393**(June), 440–442.
- [47] STRONA, G., NAPPO, D., BOCCACCI, F., FATTORINI, S. & SAN-MIGUEL-AYANZ, J. (2014) A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications*, **5**, 4114.
- [48] UGANDER, J., BACKSTROM, L., MARLOW, C. & KLEINBERG, J. (2012) Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, **109**(16), 5962–5966.
- [49] VERHELST, N. D. (2008) An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika*, **73**(4), 705.
- [50] VIGER, F. & LATAPY, M. (2005) Efficient and simple generation of random simple connected graphs with prescribed degree sequence. in *International Computing and Combinatorics Conference*, pp. 440–449. Springer.
- [51] YIN, H., BENSON, A. R. & LESKOVEC, J. (2018) Higher-order clustering in networks. *Physical Review E*, **97**(5).
- [52] YOUN, H., STRUMSKY, D., BETTENCOURT, L. M. A. & LOBO, J. (2015) Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society, Interface*, **12**(106), 1–8.
- [53] YOUNG, J. G., PETRI, G., VACCARINO, F. & PATANIA, A. (2017) Construction of and efficient sampling from the simplicial configuration model. *Physical Review E*, **96**(3), 1–6.