EDGE EXCHANGEABILITY: A NEW FOUNDATION FOR MODELING NETWORK DATA

HARRY CRANE AND WALTER DEMPSEY

Abstract. We provide a non-technical description and motivation for the edge exchangeable framework of network modeling. The discussion here complements our prior work, in which we introduced and developed the basic theory and philosophy of edge exchangeability in detail, and is intended to provide further details on our seminar presentations on the topic.

The need for edge exchangeability as a new foundation for network modeling arises out of the observation that traditional exchangeable models for vertex labeled graphs cannot replicate the large sample behaviors of sparsity and power law degree distributions observed in many network datasets. Beyond addressing this basic issue, the edge exchangeable framework promotes a mindset that better explains key features of network analysis. It also clarifies why the conventional approach is untenable for networks constructed from processes of interactions.

1. Introduction

We first introduced the concept of *edge exchangeability* in a series of articles [11, Section 6.3] and [9, 12] and later expounded the details in [13, 14]. Edge exchangeability offers a new framework for modeling network data which accords with basic logical and empirical considerations that mainstream approaches fail to acknowledge. Here we emphasize the statistical motivation and practical implications of the edge exchangeable framework and explain the need for a new foundation for network modeling in the first place.

The following commentary complements many colloquium presentations we have given on this material since 2015; see, e.g., [7]. The discussion below is intended as an accessible companion to the more technical developments in [12, 13, 14].

1.1. Principles of network modeling.

The need for a new foundation of network modeling arises out of a well known discrepancy between the behavior exhibited by real world network data and networks described by leading statistical network models, such as graphon models [5, 20], exponential random graph models [15], and the stochastic blockmodel [16]. On the one hand are the structural features of many real world networks, which tend to be *sparse* and have *power law degree distributions* [1, 2, 4, 21]. Debate lingers about the genesis of these properties, whether they are caused by the dynamics of network formation, as in preferential attachment-type models [4, 6], or bias inherent in network sampling, as argued in [19, 22]. Which of these claims, if any, is correct varies among applications, and in many situations it seems likely that both elements contribute to the observed structure. With this in mind, we devised our general framework for network modeling [12] to address the

Date: September 27, 2016.

Key words and phrases. edge exchangeability; Hollywood model; network data; sparse network; power law distribution; exchangeable random graph.

H. Crane is partially supported by NSF grants CNS-1523785 and CAREER DMS-1554092.

Dataset	vertices	edges
Actor collaborations	actors	movies
Enron email corpus	employees	emails
Karate club dataset	club members	social interactions
Wikipedia voting	Wikipedia admin.	votes
US Airport	airports	flights
Scientific collaborations	scientists	articles
UC Irvine online community	members	online messages
Political blogs	Websites	hyperlinks

TABLE 1. List of common network datasets along with a description of the entities comprising their vertices and edges. For example, the actor collaboration dataset records the set of actors (vertices) in a collection of movies (edges). The Enron email corpus is a collection of email correspondences (edges) among employees in the Enron Corporation (vertices).

effect of the data generating process and sampling mechanism at play in network formation.

Still, many aspects of data analysis result from decisions that are external to both the data generating process and sampling scheme. If not accounted for, these decisions, such as choice of sample size, experimental design, and data representation, can have a drastic effect on statistical inferences and, in the worst case, lead to bogus conclusions. One immediate feature of so-called *network data* that is universally overlooked is the step taken in representing the data by a graph G with vertex set V and edge set $E \subseteq V \times V$.

Importantly, the representation of network datasets as a graphical structure, usually as a graph $G = (V, E)$, is a modeling decision in its own right. Most network datasets, such as those in Table 1 below, do not arrive in the form of Figures 1 and 2. The effect of this choice in representing the data have gone overlooked in the vast literature on networks.

Buried in this seemingly harmless step are several critical assumptions: the representation by a graph captures the relevant information in the data; the inherent labeling assigned to elements by V does not affect inference; and, most subtly, the act of labeling vertices both implicitly identifies the vertices as sampling units and asserts that vertices can be identified independently of their relationships (via the network structure) to other vertices. Appreciating the implications of this latter assumption is crucial to the mindset of edge exchangeability put forth below.

2. NETWORK MODELING

2.1. **Common network datasets.** Table 1 lists some common network datasets. The actors collaboration network, for example, records interactions among actors based on their collaboration in movies. Each edge represents a different movie, with adjacent vertices corresponding to the set of actors who play a role in that movie. In the Enron email corpus, edges correspond to emails exchanged between employees in the Enron Corporation.

In these and most other examples in Table 1, edges can involve more than two vertices—for example, movies generally involve more than two actors and emails within a company are often exchanged among several recipients—and the same

collection of vertices can occur in multiple edges—that is, the same actors can, and often do, appear together in multiple movies and the same individuals often exchange multiple emails back and forth. Most network models, on the other hand, are specifically tailored to a $\{0,1\}$-valued relation among pairs of vertices, a structure satisfied in remarkably few real applications.

The practice of fitting these network models to the data, therefore, involves a further step in which the above structural information is destroyed by compressing hyperedges among multiple vertices, say the three-way interaction $\{a,b,c\}$, to a collection of all possible pairwise interactions, $\{a,b\},\{b,c\},\{a,c\}$, and removing edge multiplicities by thresholding to 0 (absent) or 1 (present). This action introduces a further assumption that the thresholding operation preserves any relevant features in the data. There seems an unreasonable expectation that whatever relevant features of the data drive the formation of interactions will still be apparent in the post-processed network, but such sensitivity analyses are rarely performed. We highlight some consequences of this presumption in Section 3.4.

2.2. **The edge exchangeable mindset.** Most network datasets are constructed in a way that depends on the network structure. In most applications, there is no *a priori* way to identify the population of vertices from which the network is sampled except by reference to the basic defining characteristics of the network itself. Think, for example, of the network corresponding to collaborations among scientists. The population of scientists comprising the vertices of this network corresponds precisely to those scientists who are listed as authors on an article in the database of publications. The population does not consist of all scientists, or even all physicists, all chemists, etc. Rather, the population is defined by self-reference to the process of interactions, in this case publications, that generates the network structure.

When studying the structure of such networks, it is the way in which the interactions come together—that is, interactions between interactions—that is of main interest. The identities of the vertices play no role and convey no additional information. It is appropriate to de-identify vertices by either removing their labels or choosing not to label them in the first place. This decision reflects the fact that the observed authors are generated by the process itself and, thus, have no identity beyond their interactions with other authors, as encoded by the network structure. On the flip side, the interactions comprising the edges can be identified independently of the rest of the network, as when edges represent email exchanges or movie collaborations, making the act of labeling edges natural for representing these datasets.

In these and the other examples in Table 1, network growth is driven by new interactions, that is, edges, among the vertices. These preliminary, yet crucial, observations motivate our development of edge exchangeability, which not only identifies the edges as the statistical units but recognizes their precedence in determining the identities of the vertices in the network.

2.3. **Conventional approach.** Before delving into the details of edge exchangeability in Section 3, we first highlight the main issues with the conventional approach to network modeling as it is mainly presented in the current literature.

The examples in Table 1 are commonly regarded as "network datasets". We stress that *network data*, in the sense we intend here, is a graphical representation of relational or interaction data from a sample, as in Figure 1. Given the varied applications from which network data arise, we leave this definition purposely vague. It is assumed that the complex structure produced by these interactions

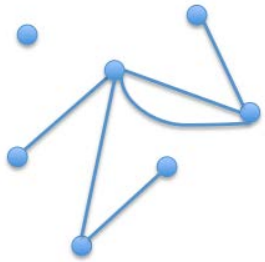

FIGURE 1. Visual representation of the sufficient information contained in network data. We assume no additional information is carried by vertex or edge labels.

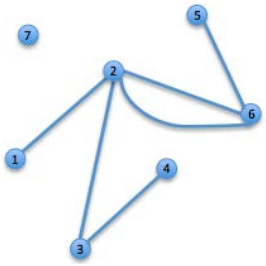

FIGURE 2. Assignment of vertex labels to the network data in Figure 1. The representation by a vertex labeled graph is common as it puts network data in a familiar context of random graphs, which are easy to store computationally and for which there is a well developed technical apparatus.

contains critical information about the underlying process of interest. We assume no further information is carried by covariates or labels on the vertices or edges.

Though the relevant information in the data is given by the structure in Figure 1, it is common practice to represent the interaction data as a graph $G = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$ as in Figure 2. The vertices correspond to elements of the population with each edge $(i, j) \in E$ representing a connection, interaction, or relationship between elements corresponding to vertices i and j. In this setting, a network model is a family of probability distributions on graphs with labeled vertices.

We have already highlighted several drawbacks to this approach. Namely, the graphical representation $G = (V, E)$ tacitly assumes that each interaction is a binary measurement, either $(i, j) \in E$ or $(i, j) \notin E$, or that a binary measurement adequately captures the information in the interactions via thresholding. This setup also assumes that multiway interactions can be flattened to a collection of pairwise interactions. For example, the interactions in the actors and Enron networks, and several others, need not be restricted to a pairwise interaction involving exactly two vertices. It is, in fact, rare for a movie cast to consist of exactly two actors, and many emails exchanged within a company involve a list of several recipients. Few existing models are flexible enough to directly handle this situation.

Aside from these apparent differences between the real data and the chosen representation of the data, basic logical and empirical assumptions are crucial to guard against spurious inferences, as we discuss in the coming section.

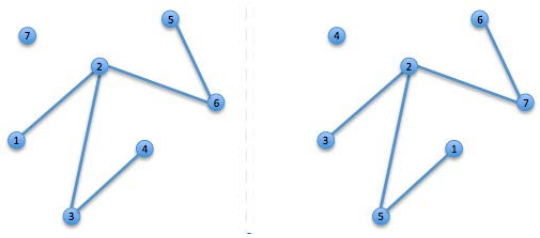

FIGURE 3. Two different assignments of vertex labels to the same network dataset. Under an exchangeable model, both of these observations are assigned equal probability.

2.4. **Logical and empirical properties.** We seek models that exhibit a bare minimum of logical and empirical properties necessary to capture the observed features in the data and permit sensible statistical inferences without falling prey to arbitrariness introduced during data analysis. For our purposes here, the basic logical properties are consistency of finite sample distributions and invariance of the model under arbitrary relabeling of the vertices. In the context of models for vertex labeled graphs, this latter invariance leads naturally to an *exchangeable model*, that is, one which assigns the same probability to any two graphs that are equivalent up to relabeling, as in Figure 3.

For empirical properties, we stick to the basic properties of sparsity and power law that have driven much of the interest in network science since the late 1990s, e.g., [1, 2, 4, 21]. These properties assume a sequence of networks $G_1, G_2, \ldots$ of growing sample size. In the present setting of vertex labeled graphs, we assume $G_n = ([n], E_n)$ has vertex set $[n]$ obtained by sampling n individuals from the population and observing the interactions among them. We, therefore, regard the number of vertices $v(G_n) = n$ as the sample size and define $(G_n)_{n\geq 1}$ to be *sparse* if

$$\limsup_{n\to\infty} \frac{e(G_n)}{v(G_n)^2} = 0, \tag{1}$$

where $e(G_n)$ is the number of edges in G_n. As a technical point, we take the limit superior in (1), instead of the limit, because the limit need not exist for a given sequence of graphs.

An intuitive way to understand (1) is to assume we take a simple random sample S of finitely many vertices and we look at the subgraph $G_n|_S$ induced on those vertices as $n \to \infty$. A consequence of sparsity is that the sequence $(G_n|_S)_{n\geq 1}$ converges to the empty graph with probability 1 as $n \to \infty$.

The power law is a more specific statement about the degree distribution. In a graph $G = (V, E)$, the *degree* of vertex $i \in V$ is the number of edges $(j, j') \in E$ for which $i \in \{j, j'\}$; it is the number of edges incident to i. Writing $p_{k,n}$ to denote the proportion of vertices in G_n with degree $k \geq 1$, we say $(G_n)_{n\geq 1}$ exhibits *power law degree distribution* with exponent $\gamma > 1$ if

$$p_{k,n} \sim k^{-\gamma} \quad \text{for all large } k \geq 1 \quad \text{as } n \to \infty. \tag{2}$$

The relationship in (2) is often demonstrated empirically by a negative linear relationship on the log-log scale, as in Figure 4.

We stress here that both sparsity and power law degree distribution are asymptotic properties and, therefore, cannot be regarded as truly empirical features of the data. Nevertheless, it is common practice in the literature to use the heuristic that a network is *sparse* if it has few edges relative to the number of vertices. This heuristic

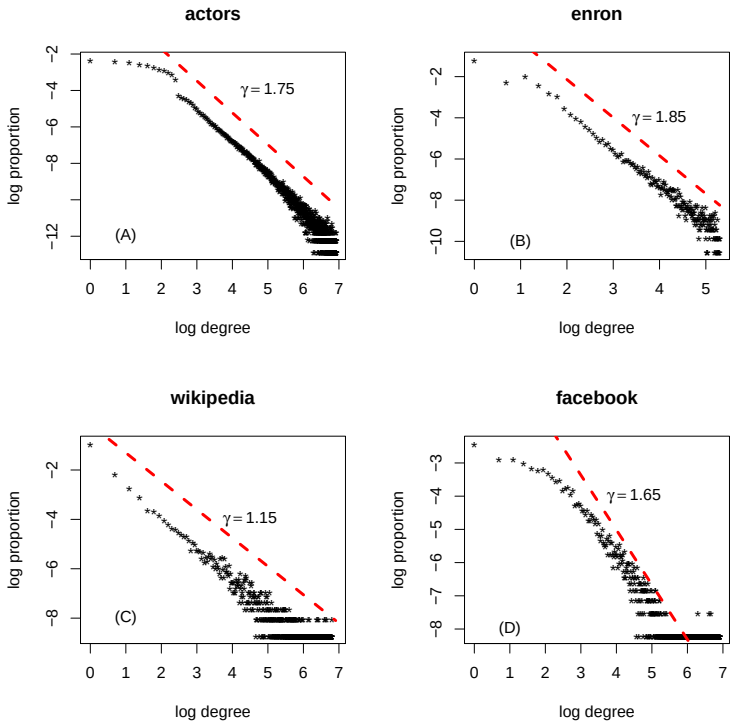

FIGURE 4. Degree distribution on the log-log scale for actors, Enron, Wikipedia, and Facebook networks. The dashed line is the power law exponent estimated by eye and may not agree with estimates given by formal calculations.

is incorrect and ought to be avoided. Instead, we take the assertions of sparsity and power law themselves as assumptions about the way in which the network structure varies with sample size. Such properties can never be confirmed from any finite amount of data, but they can nevertheless be verified, at least informally, from empirical observation. This latter aspect distinguishes these properties from the assumptions of exchangeability and consistency of finite sample distributions, which are basic assumptions of the chosen model, rather than properties of the data, and therefore cannot be validated empirically.

2.5. **Fundamental problem.** Neither sparsity nor power law can hold for a sequence $(G_n)_{n\geq 1}$ drawn by sampling from an infinite, exchangeable population graph G, unless the population graph is itself empty, that is, has no edges or sampling is performed in a biased way. This observation follows immediately by the Aldous–Hoover theorem for exchangeable random arrays [3, 17], which is closely tied to the now popular class of *graphon models* [20].

The observation presents a fundamental problem in statistical network modeling: the basic considerations noted above—exchangeability and consistency of finite sample distributions on the logical side and sparsity and power law on the empirical side—cannot be accommodated by any exchangeable model for network data *when the data are represented as a graph with labeled vertices* as in Figure 2.

With this observation, it is curious that graphons have caught on as they have in the theoretical statistics literature on networks as these basic observations call

into question, and seem to negate, their practical value. We have not heard a cogent argument in favor of using graphon models for real network data.

It is widely agreed that the Erdős–Rényi model, which assumes each edge is present independently with a fixed probability $p \in (0,1)$, is untenable as a practical model for network data. But, in a sense made precise by the Aldous–Hoover theorem, the class of graphon models refines the Erdős–Rényi construction while preserving its distinguishing features. Instead of modeling each edge as independent, identically distributed draws, the Aldous–Hoover theorem models each edge as a conditionally independent draw given random vertex effects. The net effect is a class of models which is unable to replicate the most basic features observed in network data, in much the same way and for the same reasons as the Erdős–Rényi model.

Observation 2.1. *The conventional approach to network modeling cannot simultaneously account for basic logical and empirical considerations.*

The current situation, therefore, is not hospitable to principled statistical modeling. This series of observations, at first glance, is rather vexing, as there appears nothing fallacious in labeling vertices of the network and then removing the effect of this labeling by choosing an exchangeable model. However, as we discover below, the effect of labeling vertices is not fully eliminated by assuming an exchangeable model. Assigning labels has a lasting effect that, once introduced, cannot be undone. Our main observation is that the majority of network datasets mentioned at the outset ought not be represented as a vertex labeled graph in the first place.

3. Edge exchangeable models

To realize the consequences of our observation in Section 2.2 that edges comprise the statistical units in many common network datasets, take the concrete example of the actors collaboration network. Sampling this network does not proceed by taking a simple random sample of people and asking in which movies they have acted. Such an approach would, with high probability, result in an empty graph. A more reasonable approach is to sample movies from a database, such as the Internet Movie Database (IMDB), and observe the corresponding sets of actors in the chosen movies. Sampling movies uniformly without replacement from the IMDB results in a network with edges labeled according to the order in which they were sampled.

There are two important aspects of this mode of sampling. First is that the sample is driven by a process on the movies, which correspond to the edges in the corresponding network representation. The sampled actors are incidental to the edge sampling scheme and cannot be identified without reference to the interactions (movies) in which they have acted. Indeed it is impossible to identify someone as an actor without also identifying a movie in which he/she has acted. The labels can be assigned only after movies have been sampled; we cannot label unsampled vertices prior to sampling them. The act of labeling the vertices in the graph in Figure 1 is, at best, misleading and, at worst, incorrect since it suggests that unobserved vertices can be identified independently of their interactions. In fact, unobserved vertices cannot be identified without being sampled.

These observations lead to a simple formalism for *interaction data*, which we define as a process $(E_i)_{i\in\mathbb{N}}$ of subsets of a population $\mathcal{P}$. In the following example, each $E_i \subset \mathcal{P}$ need not be of size 2, but we restrict to this case for simplicity.

Example 3.1. *Suppose a population* $\mathcal{P} = \{1,\ldots,10\}$ *and a sequence of interactions*

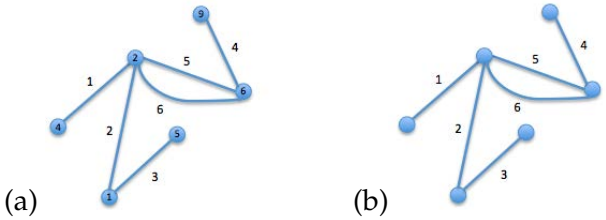

FIGURE 5. (a) Vertex and edge labeled graph obtained by sequence of interactions $E_1 = \{2,4\}$, $E_2 = \{1,2\}$, $E_3 = \{1,5\}$, $E_4 = \{6,9\}$, $E_5 = \{2,6\}$, $E_6 = \{2,6\}$. (b) Edge labeled graph obtained by removing vertex labels from graph in Panel (a).

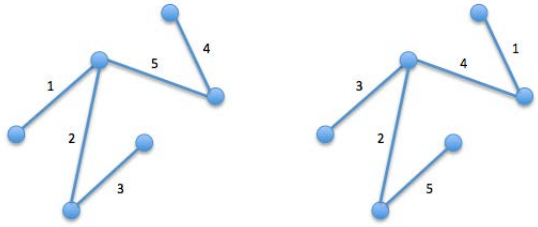

FIGURE 6. Relabeling of two edge labeled graphs. An edge exchangeable models assigns equal probability to both outcomes.

- $E_1 = \{2,4\}$, $E_2 = \{1,2\}$, $E_3 = \{1,5\}$, $E_4 = \{6,9\}$, $E_5 = \{2,6\}$, $E_6 = \{2,6\}$

We can record these interactions, without any loss of generality, as in the vertex and edge labeled graph of Figure 5(a). However, as we mentioned above for the actors and Enron networks, the vertex labels serve only to keep track of vertices during the course of network formation. Vertex labels should not be construed as equipping sampled individuals with an identity beyond their interactions with other vertices in the network. We, therefore, remove vertex labels to arrive at the representation of the relevant structure in the network by an edge labeled graph as in Figure 5(b).

As presented in Example 3.1, the edge labels in Figure 5(b) correspond to a temporal ordering in which the interactions occurred. In practice, however, this temporal ordering may often be ignored for a few possible reasons. Most commonly, as in the actors collaboration dataset, the temporal order of movies is known to exist but is not recorded in the data. In this case, the observation is an unlabeled structure as in Figure 1, so that modeling the data as an edge labeled graph entails appending edge labels arbitrarily to the observed structure. But even if the ordering is observed, as it would be if we sampled movies ourselves from the IMDB, it is sometimes reasonable to treat the data symmetrically with respect to the observed ordering, as it would be if we sampled movies without replacement from the IMDB.

In either case, the principle of exchangeability from Section 2.3 leads naturally to the concept of *edge exchangeability*, by which any two graphs which are equivalent up to relabeling of edges are assigned equal probability. Figure 6 demonstrates two such graphs, which is sufficient as a definition of edge exchangeability for the purposes of this discussion. See [13] for a complete definition and further discussion.

As a logical principle, edge exchangeability is more natural than the conventional approach to network modeling with random graph models. Edge exchangeability,

it turns out, also resolves the main empirical issues of incorporating sparsity and power law degree distribution into a principled statistical model. We discuss these below.

Given these observations, it is perhaps surprising that we first introduced edge exchangeability only recently in the context of a broader discussion on statistical network modeling; see [11, Section 6.3]. An explanation of this, we believe, is that the consequences of representing network data as a graph with labeled vertices are not understood and have never been pointed out before. In fact, in most outlets *network data* is treated as interchangeable with *graphical data*.

3.1. **Vertices arrive in size-biased order.** One immediate consequence of the edge-centric view taken by the interaction process of Example 3.1 is that the sample size is defined as the number of interactions (or edges) in the observed network. The number of vertices is a random variable depending on the data generating process and the sample size of edges.

In [14], we proved a general representation theorem for edge exchangeable networks, and more general *relationally exchangeable* structures, which sheds light on the process by which vertices arrive. Avoiding technicalities here, we consider as a special case an array $(f_{\{i,j\}})_{j\geq i\geq 1}$ for which each $f_{\{i,j\}} \geq 0$ and $\sum_{j\geq i\geq 1} f_{\{i,j\}} = 1$; see [14] for the general case.

Let $\mathcal{F}$ denote the space of all such arrays. We construct an edge exchangeable graph by first specifying a probability measure ϕ on $\mathcal{F}$ and selecting $f = (f_{\{i,j\}})_{j\geq i\geq 1}$ randomly according to ϕ. Given f, we sample interactions $X_1, X_2, \ldots$ conditionally independently and identically distributed (i.i.d.) according to

$$\mathbb{P}(X_n = \{i,j\} \mid f) = f_{\{i,j\}}, \quad j \geq i \geq 1.$$

From $X_1, X_2, \ldots$, we construct first the vertex-edge labeled graph induced by these sets, as in Figure 5(a), and then remove vertex labels to obtain a random edge labeled graph as in Figure 5(b). The resulting graph is edge exchangeable. By Theorem 3.4 of [14], every edge exchangeable network admits such a construction for some ϕ.[1]

While this representation may, on its own, be of use for modeling edge exchangeable networks using techniques from completely random measures and Bayesian nonparametrics, it offers an immediate practical insight. Notice the effect of the construction of $X_1, X_2, \ldots$ as i.i.d. from f on the arrival of vertices. For each $i \geq 1$, let $f_{i\bullet} = \sum_{j\geq 1} f_{\{i,j\}}$ be the sum of the weights over all edges incident to i. Then $f_{i\bullet}$ is the probability that vertex i is contained in any given edge $X_1, X_2, \ldots$, leading to the following observation.

Observation 3.2. *Vertices arrive in an edge exchangeable graph in* size-biased order *according to their expected relative degrees* $f_{i\bullet}$.

Though not usually phrased or regarded in this way, the more common assumption of vertex exchangeability implicitly assumes that new vertices arrive in exchangeable random order. Observation 3.2 establishes that the behavior of edge exchangeable networks is inconsistent with this assumption, explaining why the usual assumption of vertex exchangeability, which at first glance appears reasonable and innocuous, is expressly violated by the assumptions of most network models.

[1] In fact, Theorem 3.4 of [14] states the stronger result that every edge exchangeable random graph has a unique representation in terms of some canonical measure ϕ, but we do not discuss that generality here.

We note that size-biased arrival of vertices is critical to other models for sparse, power law network structures, notably the preferential attachment-type models [4, 6]. We observe the same empirical behavior for certain edge exchangeable models.

3.2. **Sparsity and power law in the Hollywood model.** Let $\nu = \{\nu_k\}_{k\geq 1}$ be a probability distribution on the positive integers and let (α, θ) satisfy either $0 < \alpha < 1$ and $\theta > -\alpha$ or $\alpha < 0$ and $\theta = -k\alpha$ for some $k = 1, 2, \ldots$. We generate a sequence $(Y_n)_{n\geq 0}$ of edge labeled networks, with each Y_n having $n \geq 1$ edges, as follows.

Write $v(Y_n)$ to denote the number of vertices in Y_n and $e(Y_n)$ to denote the number of edges in Y_n. We start with Y_0 having $v(Y_0) = e(Y_0) = 0$. In the intended semantic interpretation as a process of movie formation, every edge in Y_n corresponds to a movie, with the vertices incident to the edge labeled i in Y_n corresponding to the actors who play a role in movie $i = 1, \ldots, n$. We stress, however, that the application to movie formation is not central to the Hollywood model.

Given Y_{n-1}, for $n \geq 1$, we choose the number of available roles in the next movie independently according to $K_n \sim \nu$. Given $K_n = k$, we choose the k actors in order of their prominence, first filling the lead role, then the second lead role, and so on until all k roles are filled. Let $N_n(j)$ be the number of unique actors seen in all the movies through the $(j-1)$st actor cast in movie n. (Thus, $N_n(1)$ is the number of unique actors appearing in movies $1, \ldots, n-1$.) For $j = 1, \ldots, k$, we label the actors arbitrarily $1, \ldots, N_n(j)$ and write $D_n(i, j)$ to denote the number of roles for which the actor labeled i has been cast up to and including the $(j-1)$st role of movie n. (Note that an actor may play more than one role in a given movie.) The actor $v_n(j)$ cast in the jth lead role of movie n is chosen randomly among the actors labeled $1, \ldots, N_n(j)$ and a previously unseen actor, labeled $N_n(j) + 1$, according to

$$\text{pr}(v_n(j) = i) \propto \begin{cases} D_n(i, j) - \alpha, & i = 1, \ldots, N_n(j), \\ \theta + \alpha N_n(j), & i = N_n(j) + 1. \end{cases} \tag{3}$$

We continue to update according to (3) until all k roles of movie n have been filled.

Example 3.3. *Let ν be a probability distribution on the positive integers. Suppose $K_1, K_2, \ldots$ are i.i.d. from $\nu = \{\nu_k\}_{k\geq 1}$, $0 < \alpha < 1$, and $\theta > -\alpha$.*

Then for $(K_1, K_2, \ldots) = (3, 2, 4, \ldots)$:

- $E_1 = (1, 2, 1)$ *with probability*

$$\frac{\theta}{\theta} \times \frac{\theta + \alpha}{\theta + 1} \times \frac{1 - \alpha}{\theta + 2}$$

- $E_2 = (3, 2)$ *with probabilty*

$$\frac{\theta + 2\alpha}{\theta + 3} \times \frac{1 - \alpha}{\theta + 4}$$

- $E_3 = (1, 4, 3, 5)$ *with probability*

$$\frac{2 - \alpha}{\theta + 5} \times \frac{\theta + 3\alpha}{\theta + 6} \times \frac{1 - \alpha}{\theta + 7} \times \frac{\theta + 4\alpha}{\theta + 8}$$

The probability of (E_1, E_2, E_3) here is given by

$$\nu_3 \times \nu_2 \times \nu_4 \times \alpha^5 \frac{(\theta/\alpha)(\theta/\alpha + 1) \cdots (\theta/\alpha + 4)}{\theta(\theta + 1) \cdots (\theta + 8)} (1 - \alpha)^3 (1 - \alpha + 1).$$

We give a general expression in (4).

We call the sequence $(Y_n)_{n\geq 1}$ of networks constructed this way the *Hollywood process* with parameter (α, θ, ν). By its sequential construction, the process determines a family of distributions, called the *Hollywood model*, for network data with edges labeled in $\mathbb{N}$. We express the distribution of Y_n, for each $n \geq 1$, in closed form by

$$\text{(4)}\quad \text{pr}(Y_n = E; \alpha, \theta, \nu) = \left[\prod_{k\geq 1} \nu_k^{M_k(E)}\right] \alpha^{v(E)} \frac{(\theta/\alpha)^{\uparrow v(E)}}{\theta^{\uparrow m_n(E)}} \prod_{k=2}^{\infty} \exp\{N_k(E) \log((1-\alpha)^{\uparrow(k-1)})\},$$

where E is any edge labeled network with n oriented edges, $v(E)$ is the number of nonisolated vertices in E, $(N_k(E))_{k\geq 0}$ gives the number of vertices with degree k for each $k \geq 0$, $M_k(E)$ is the number of k-ary edges in E, $m_n(E) = \sum_{k\geq 1} k M_k(E)$ is the total degree of E, and $x^{\uparrow j} = x(x+1)\cdots(x+j-1)$ is the ascending factorial function.

Though presented in the context of the actors network, the Hollywood model is suited to any conceivable network dataset that arises from a process of interactions. The special case of network data with binary edges is easily handled by setting $\nu_2 = 1$. We discuss further practical aspects of the Hollywood model in [13].

3.3. **Vertex components models.** The above Hollywood model was first introduced under the heading *Poisson–Dirichlet model* in [11, Section 6.3]. This previous representation is a special case of the vertex components processes we now describe.

Crane [8] previously noted a connection between the Hollywood model and the two parameter Poisson–Dirichlet distribution as follows. Given (α, θ) in the parameter space of the Hollywood model, we let $W = (W_1, W_2, \ldots)$ be a random draw from the Poisson–Dirichlet distribution with parameter (α, θ) and define $f_{\{i,j\}} = W_i W_j$ for each $j \geq i \geq 1$. This construction determines a probability measure $\phi_{\alpha,\theta}$ on $\mathcal{F}$ as described in Section 3.1 above.

More generally, this construction of $f = (f_{\{i,j\}})_{j\geq i\geq 1}$ by $f_{\{i,j\}} \propto W_i W_j$ for some sequence of random variables $(W_1, W_2, \ldots)$ with $\sum_{i\geq 1} W_i < \infty$ is a special case of the *vertex components* process first introduced in [11, Section 6.3]. The vertex components models are a tractable class of nonparametric models for edge exchangeable networks which may have additional practical benefits beyond the Hollywood model.

From this connection between the Hollywood model and the vertex components model driven by the Poisson–Dirichlet distribution, we can deduce, see [13, Section 5.3], that the Hollywood process $(Y_n)_{n\geq 1}$ exhibits power law degree distribution with exponent $\gamma = \alpha + 1$ when $0 < \alpha < 1$ and $\theta > -\alpha$. Furthermore, $(Y_n)_{n\geq 1}$ is sparse provided $1/\mu < \alpha < 1$, where $\mu = \sum_{k\geq 1} k\nu_k$ is the mean interaction size in the general Hollywood model with parameter (α, θ, ν).

3.4. **Unintended consequences of thresholding.** We previously alluded to the consequences of thresholding multiple edges to obtain an ordinary graph. Though unnecessary and detrimental to inference, the act of thresholding is very common in practice as alluded above. The practical reason for projecting multiple edges seems to be closely tied to the fact that most network models are unable to accommodate the natural occurrence of multiple edges. The Hollywood model does not suffer from these issues as it directly models networks with multiple edges and with interactions involving more than two vertices, as is common in all of the examples of Table 1.

Theorem 5.4 of [13] demonstrates an immediate consequence of removing multiple edges from a multigraph constructed from the Hollywood model: the projection of $(Y_n)_{n\geq 1}$ to a sequence of simple graphs by removing multiple edges is sparse for

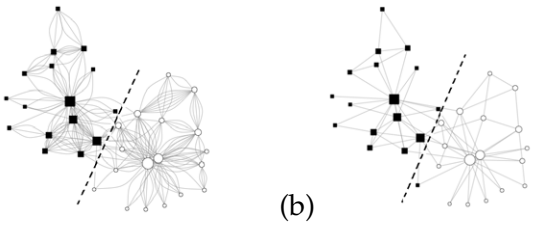

FIGURE 7. (Left) Interaction network of Zachary karate club. (Right) Projection by removing edge multiplicities. In both pictures, the dashed line separates vertices into the true communities, as dictated by Zachary [23]. The color of the vertices, as black or white, shows the classification given by either a simple method that accounts for multiple edges (in (a)) or more complicated approaches, such as degree-corrected stochastic blockmodels, which ignore multiplicities (in (b)). The classification in (a) coincides with Zachary's analysis while the classification in (b) does not.

all $0 < \alpha < 1$. Therefore, thresholding not only throws away data unnecessarily but also fundamentally changes the asymptotic behavior of the data.

One place where thresholding is pronounced is in community detection applications, as in Figure 7. Figure 7(a) shows the complete karate club dataset as originally recorded in [23]. That dataset records the number of social interactions between each pair of the 34 members of a university karate club. The karate club dataset is a canonical example in network community detection because of the known separation of club members according to their allegiance to one of the club's two leaders. However, most community detection techniques we know of have been illustrated on the karate club network after projecting the multiple interactions to a single edge, as shown in Figure 7(b).

It is common, e.g., in [18], for the most successful methods to properly classify all but one of the karate club members, as shown in Figure 7. However, the community structure imposed by social interactions among members of the club is related directly to the multigraph in Figure 7(a). There is no logical reason to expect the same community structure to persist after arbitrarily projecting multiple edges. In fact the leading community detection methods, e.g., [5, 18], consistently misclassify one of the members.

From a logical point of view, this ought to be expected, since the act of removing multiple edges offers no guarantee to preserve the fundamental structure of the data. In fact, we have shown in other work [10] that all vertices can be correctly classified with a very simple approach if only the complete dataset with multiple edges is analyzed.

4. Concluding remarks

The mindset of edge exchangeability described here is the result of a logical flow from data arising by an interaction process, to its representation by an edge labeled graph, to the notion of edge exchangeability. From the discussion, edge exchangeability offers an alternative to the current suite of network models, such as graphons, exponential random graph models, and stochastic blockmodels, which are unable to explain the most basic properties of network data. The fact that edge exchangeable networks replicate these basic features of network data is merely

empirical justification that the framework may be viable for the many intended applications in network science.

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. *Proceedings of the 6th European Symposium on Algorithms*, pages 332–343, 1998.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 171–180, New York, 2000. ACM Press.

[3] D. J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] P. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21068–21073, 2009.

[6] F. Chung and L. Lu. *Complex graphs and networks*, volume 107 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2006.

[7] H. Crane. Edge exchangeability: a new foundation for modeling network data. *Presentation at Isaac Newton Institute for Mathematical Sciences*, July 25, 2016. Available online at www.newton.ac.uk/seminar/20160725133014001/ Also available at stat.rutgers.edu/home/hcrane/talks/.

[8] H. Crane. Rejoinder: The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1):37–39, 2016.

[9] H. Crane and W. Dempsey. Atypical scaling behavior persists in real world interaction networks. arXiv:1509.08184, 2015.

[10] H. Crane and W. Dempsey. Community detection for interaction networks. arXiv:1509.09254, 2015.

[11] H. Crane and W. Dempsey. A framework for statistical network modeling. First version, arXiv:1509.08185v1, 2015.

[12] H. Crane and W. Dempsey. A framework for statistical network modeling. Second version, arXiv:1509.08185, 2015.

[13] H. Crane and W. Dempsey. Edge exchangeable models for network data. arXiv:1603.04571, 2016.

[14] H. Crane and W. Dempsey. Relational exchangeability. arXiv:1607.06762, 2016.

[15] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[16] P. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, pages 33–65, 1981.

[17] D. Hoover. Relations on Probability Spaces and Arrays of Random Variables. *Preprint, Institute for Advanced Studies*, 1979.

[18] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011.

[19] S. H. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.

[20] L. Lovász and B. Szegedy. Limits of dense graph sequences. *J. Comb. Th. B*, 96:933–957, 2006.

[21] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[22] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the Internet: a source of enormous confusion and great potential. *Notices Amer. Math. Soc.*, 56(5):586–599, 2009.

[23] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

Department of Statistics & Biostatistics, Rutgers University, 110 Frelinghuysen Avenue, Piscataway, NJ 08854, USA

E-mail address: hcrane@stat.rutgers.edu

URL: http://stat.rutgers.edu/home/hcrane

Department of Statistics, University of Michigan, 1085 S. University Ave, Ann Arbor, MI 48109, USA

E-mail address: wdem@umich.edu