Statistical Science
2021, Vol. 36, No. 1, 51–67
https://doi.org/10.1214/19-STS759

A Statistical Framework for Modern Network Science

Harry Crane and Walter Dempsey

Abstract. We discuss how sampling design, units, the observation mechanism and other basic statistical notions figure into modern network data analysis. These considerations pose several new challenges that cannot be adequately addressed by merely extending or generalizing classical methods. Such challenges stem from fundamental differences between the domains in which network data emerge and those for which classical tools were developed. By revisiting these basic statistical considerations, we suggest a framework in which to develop theory and methods for network analysis in a way that accounts for both conceptual and practical challenges of network science. We then discuss how some well-known model classes fit within this framework.

Key words and phrases: Network data, sparse network, scale-free network, edge exchangeable network, relational exchangeability, relative exchangeability, data generating process, network sampling.

1. INTRODUCTION

The earliest methods for network analysis were developed in the quantitative social sciences, beginning with Moreno's 1930 introduction of the sociogram [39] and continuing with the development of stochastic blockmodels (SBMs) [28] and exponential random graph models (ERGMs) [23, 29]. Since the mid-1990s, however, the focus of network analysis has shifted from social networks toward large, complex networks that emerge in applications across the social, biological and physical sciences. We aim here to clarify a few important aspects of statistical analysis in this new landscape of network science. Our discussion culminates in a proposed framework for network modeling that demonstrates how classical statistical concepts such as sampling design and observational units figure into theoretical and methodological developments. A key element of our discussion is how statistical methods can be built around models that capture known empirical behavior in observed network data while accounting for basic tenets of sensible statistical inference and computational limitations of complex data structures. The discussion subsumes classical network models, such as SBMs, ERGMs and graphon models, as well as some more recent proposals (e.g., the Crane–Dempsey edge exchangeable framework [16] and the Caron–Fox model based on completely random measures [8]) that seek to move beyond the limitations of these classical models.

After introducing our proposed modeling framework in Section 2, we move on to the simple setting of binary relations observed for an *entire* population (Section 3), so that the challenge of accounting for the sampling scheme does not arise. We then consider the more common situation in which the observed network is sampled from a larger population network. Sampling designs are detailed in Section 4, followed by a discussion of model coherence in Section 5. Section 6 discusses the interplay of model coherence and inferential tasks, clarifying its role in both in- and out-of-sample inference. Against the backdrop of the modeling framework laid out in Sections 2, 5 and 6, we then discuss the specific model classes of vertex exchangeable models (Section 7), relatively exchangeable models (Section 8.3), edge exchangeable models (Section 9) and relationally exchangeable models (Section 10).

The forthcoming sections aim to provide a broad overview of a wide swath of work scattered throughout the statistics and probability literature. We focus here on the high-level concepts behind recent developments, deferring technical details to the main references, for example, [15, 16].

Harry Crane is Associate Professor, Department of Statistics & Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854, USA (e-mail: hcrane@stat.rutgers.edu). Walter Dempsey is Assistant Professor, Department of Statistics, Harvard University One Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: wdempsey@uchicago.edu).

2. NETWORK MODELING PARADIGM

The variety of situations in which modern network data arises, for example, from fMRI images, social media interactions, paths between Internet servers and cryptocurrency transactions, necessitates a mathematical framework that allows flexibility in how to represent, model and analyze such data. With this in mind, we stress here the distinction between "network" as a scientific concept and "graph" as a mathematical structure. Whereas the concept of "network" conjures the vague but intuitive image of a complex system of interrelated entities, the mathematical notion of "graph" is defined precisely as a set of vertices V together with a binary relation $E \subseteq V \times V$. Because networks arise in a variety of situations, it may not always be appropriate or ideal to represent network data as a graph, for example, in the interaction network data studied in [16] and discussed in Section 9 below. Ultimately, the best representation of a network depends fundamentally on how the network data has been obtained, whether by sampling, a generative process, or some other means, bringing the concept of *statistical units* front and center in the development of theory and methods for network analysis.

2.1 Statistical Units for Network Data

In experimental design, the *statistical units* are defined as the "experimental units, plots or subjects" [37] of a given study, which are traditionally the smallest entities to which a treatment can be assigned. In network analysis, however, units are not entities receiving a treatment but more often are the entities comprising the network structure. Whereas the appropriate choice of unit is clear in many classical settings, for example, plots for agricultural trials or subjects for biomedical trials, it is often less clear, and varies with context, in network analysis. For example, in a high school social network obtained by observing friendships among n sampled students (as in Section 3.2), the students are the statistical units. If, however, the social network were constructed from a sample of n binary interactions among students in a high school (as in Section 9), then the interactions are the units. In both scenarios, the "social network" can be represented by a graph-like structure with students as the *vertices* and their interactions/friendships as the *edges*. But the statistical interpretation is different, with vertices as the units in the first scenario and edges as the units in the second scenario. This observation goes against conventional wisdom in the networks literature that "in most network samples, the unit of sampling is the actor or node" [25], page 7, and is a key initial observation underlying the alternative framework of edge exchangeability [16], as discussed in Section 9 below.

With the determination of units comes the related notion of sample size for network data. Though it is still common to think of network data as "a sample of size 1," the above discussion of *units* allows us to define the *sample size* of a network precisely as *the number of observed units*. To preview some upcoming examples: the sample size in the high-school friendship network considered in Section 3.2 is the number of sampled vertices, in the phone call scenario of Section 4.3 is the number of sampled phone calls (or edges), in the coauthorship scenario of Section 4.4 is the number of sampled articles (or hyperedges), and in the traceroute scenario of Section 4.5 is the number of sampled paths. See [15], Sections 3.8 and 3.9, for further discussion.

With the notions of units and sample size fixed, we often write $\mathbf{Y}_n$ to denote network data for a sample of size n, with the meaning of sample size understood implicitly by the appropriate identification of the units in the assumed situation. In the conventional representation of networks as graphs, the sample size is the number of vertices and $\mathbf{Y}_n$ is an $n \times n$ array taking values in $\{0, 1\}$. For interaction data, as studied in [16], $\mathbf{Y}_n$ is a graph with n labeled edges. The framework presented in the next section applies more generally to most current conceptions of network data, for example, as point processes (Section 8.2) or as relationally-labeled structures (Section 10).

2.2 Modeling Paradigm

Unless otherwise noted, we focus on network data $\mathbf{Y}_n$ assumed to have been obtained by sampling from a population network $\mathbf{Y}_N$ of size $N > n$. In this situation, a statistical model consists of two primary components:

(M1) The *descriptive component* articulates the sources of variation and uncertainty in the observed data by specifying, for each possible sample size n, a set of candidate distributions $\mathcal{M}_n$ for $\mathbf{Y}_n$.

(M2) The *inferential component* links the observed data $\mathbf{Y}_n$ to the unobserved population $\mathbf{Y}_N$ by a family $\mathcal{S}_{n,N}$ of candidate sampling mechanisms that describe how $\mathbf{Y}_n$ was obtained from $\mathbf{Y}_N$.

In the language of [15], Chapter 5, the inferential component in (M2) provides the context in which the analysis is to be interpreted. Though this setup accommodates much more general modeling considerations, such as networks evolving according to some generative process, we focus this paper on network models whose inferential component describes a sampling mechanism. In general, $\mathcal{S}_{n,N}$ can contain an arbitrary number of candidate sampling maps, just as $\mathcal{M}_n$ typically contains an arbitrary number of candidate probability distributions, but for simplicity we focus here on the case of a single sampling operation.

For example, the Erdős–Rényi–Gilbert distribution with parameter $0 \leq \theta \leq 1$ on $\{0, 1\}^{n \times n}$ is defined by

$$\Pr(\mathbf{Y}_n = \mathbf{y}; \theta) = \prod_{1 \leq i \neq j \leq n} \theta^{y_{ij}} (1 - \theta)^{1 - y_{ij}}, \quad \mathbf{y} \in \{0, 1\}^{n \times n}. \tag{1}$$

A model specified in terms of (M1) and (M2) above could take $\mathcal{M}_n = \{\Pr(\mathbf{Y}_n = \cdot;\theta) : 0 \le \theta \le 1\}$, that is, the set of all Erdős–Rényi–Gilbert distributions on $\{0,1\}^{n\times n}$ and, for each $N \ge n \ge 1$, $\mathcal{S}_{n,N} = \{\mathbf{S}_{n,N}\}$, where $\mathbf{S}_{n,N} : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$ is the *selection sampling map* which samples from $\mathbf{y} \in \{0,1\}^{N\times N}$ by "selecting" its first n rows and columns, that is, $\mathbf{S}_{n,N}(\mathbf{y}) = \mathbf{y}|_{[n]} = (y_{ij})_{1\le i,j\le n}$. (Note that "selection" is sometimes called "projectivity" or "restriction" by other authors.)

For another example with the same candidate distributions $\mathcal{M}_n$, take $\mathcal{S}_{n,N} = \{\Sigma_{n,N}\}$, where $\Sigma_{n,N}$ is now a random sampling operation obtained by choosing a set S of n vertices from $[N]$ uniformly at random without replacement and putting $\mathbf{Y}_n = \Sigma_{n,N}\mathbf{Y}_N = \mathbf{Y}_N|_S$, that is, the restriction of $\mathbf{Y}_N$ to this random vertex set. Klusowki and Wu [32, 33], for example, consider such random sampling operations as an inferential basis for counting motifs within a fixed network $\mathbf{y}_N$. That is, the family of distributions $\mathcal{M}_N$ consists of a single distribution P_N that assigns mass one to the network $\mathbf{Y}_N = \mathbf{y}_N$.

The two components, (M1) and (M2), lead to our definition of a *statistical network model* as a pair $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\mathcal{S}_{m,n}\}_{N\ge n\ge m\ge 1})$ of candidate distributions and subsampling mechanisms for each finite sample size. In general, each statistical model $\mathcal{M}_n$ is defined on a sample space $\mathcal{N}_n$ of "networks of size n," which is left implicit and depends on context. In the two examples above, for instance, $\mathcal{N}_n = \{0,1\}^{n\times n}$ is the set of all $n \times n$ adjacency arrays. n other contexts, $\mathcal{N}_n$ may be the set of edge-labeled graphs with edges labeled in $[n]$, as in Section 9 below, or some other natural representation of network data as the situation calls for it.

Together, components (M1) and (M2) allow the model to be interpreted both as a data generating process and as a framework for statistical inference. As a data generating process, we assume that $\mathbf{Y}_N$ is a population network generated according to one of the distributions in $\mathcal{M}_N$. From $\mathbf{Y}_N$, we observe $\mathbf{Y}_n$ by sampling according to one of the sampling schemes in $\mathcal{S}_{n,N}$. The setup affords a complementary interpretation in the inverse statistical inference problem, for which we assume the observation $\mathbf{Y}_n$ is distributed according to one of the candidates in $\mathcal{M}_n$, and that the relationship between sample and population via the sampling mechanism $\mathcal{S}_{n,N}$ provides the necessary link to draw inferences about the population generating process based on $\mathbf{Y}_n$, as in (6) below. How sampling affects inferences from sampled network data highlights the importance of the concept of *model coherence* introduced in Section 5.

3. BINARY RELATIONAL DATA

The most basic network data takes the form of a binary relation $R \subseteq [n]\times[n]$ among individuals labeled uniquely in $[n] = \{1,\dots,n\}$. For example, if $1,\dots,n$ label high school students, then $(i,j) \in R$ might indicate that 'i is friends with j' or that 'i and j are both members of the band'. This relation gives rise to an adjacency matrix $\mathbf{Y} = (Y_{ij})_{1\le i,j\le n}$ with

$$Y_{ij} = \begin{cases} 1, & (i,j) \in R, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \le i,j \le n. \tag{2}$$

3.1 Exponential Random Graph Models (ERGMs)

As an initial model for $\mathbf{Y} = (Y_{ij})_{1\le i,j\le n}$ observed for a population of n individuals (e.g., students, countries, etc.), the *dyad independence model* assumes that each *dyad* $D_{ij} = (Y_{ij}, Y_{ji})$, $i < j$, is distributed independently according to p_{ij} on $\{0,1\}\times\{0,1\}$. With

$$\rho_{ij} = \log\left(\frac{p_{ij}(0,0)p_{ij}(1,1)}{p_{ij}(0,1)p_{ij}(1,0)}\right) \quad \text{and}$$
$$\theta_{ij} = \log\big(p_{ij}(1,0)/p_{ij}(0,0)\big),$$

the distribution of $\mathbf{Y}$ can be expressed as

$$\Pr(\mathbf{Y} = \mathbf{y};\mathbf{p}) \propto \exp\left\{\sum_{1\le i<j\le n} \rho_{ij}y_{ij}y_{ji} + \sum_{1\le i\ne j\le n} \theta_{ij}y_{ij}\right\}, \tag{3}$$

for $\mathbf{y} = (y_{ij})_{1\le i,j\le n} \in \{0,1\}^{n\times n}$. Holland and Leinhardt's p_1 model [29] is a special case of the dyad independence model with

$$\rho_{ij} = \rho(1 \le i < j \le n) \quad \text{and}$$
$$\theta_{ij} = \theta + \alpha_i + \beta_j \quad (1 \le i \ne j \le n)$$

for parameters θ, ρ, $\alpha = (\alpha_i)_{1\le i\le n}$, and $\beta = (\beta_i)_{1\le i\le n}$. The resulting distribution of $\mathbf{Y}$ under the p_1 model with parameter $(\rho,\theta,\alpha,\beta)$ is

$$\begin{aligned}\Pr(\mathbf{Y} = \mathbf{y};\rho,\theta,\alpha,\beta) &= \left(\exp\left\{\rho\sum_{1\le i<j\le n} y_{ij}y_{ji} + \theta y_{\bullet\bullet} + \sum_{i=1}^n \alpha_i y_{i\bullet} + \sum_{j=1}^n \beta_j y_{\bullet j}\right\}\right) \\ &\quad \Big/\left(\prod_{1\le i<j\le n} \eta_{ij}\right),\end{aligned} \tag{4}$$

where $y_{i\bullet} = \sum_{j=1}^n y_{ij}$, $y_{\bullet j} = \sum_{i=1}^n y_{ij}$, $y_{\bullet\bullet} = \sum_{i,j=1}^n y_{ij}$ and

$$\eta_{ij} := 1 + e^{\rho+\alpha_i+\beta_j} + e^{\rho+\alpha_j+\beta_i} + e^{\rho+2\theta+\alpha_i+\alpha_j+\beta_i+\beta_j}, \quad 1 \le i < j \le n.$$

In Holland and Leinhardt's initial presentation, ρ is interpreted as *reciprocity* and α_i and β_j as *differential attractiveness*. Frank and Strauss's family of *Markov random graphs* extends the p_1 model to incorporate properties beyond reciprocity and differential attractiveness

[23]. Wasserman and Pattison [46] subsequently developed the Frank–Strauss model into the p^* *model*, now generically known as the *exponential random graph model* (ERGM). For $\theta_1, \ldots, \theta_k \in \mathbb{R}$ and statistics $T_1, \ldots, T_k : \{0,1\}^{n\times n} \to \mathbb{R}$, the *exponential random graph model* (ERGM) with (natural) parameter $\theta = (\theta_i)_{1\le i\le k}$ and (canonical) sufficient statistic $T = (T_i)_{1\le i\le k}$ assigns probabilities

$$\Pr(\mathbf{Y} = \mathbf{y}; \theta, T) = \frac{\exp\{\sum_{i=1}^k \theta_i T_i(\mathbf{y})\}}{\sum_{\mathbf{y}^*\in\{0,1\}^{n\times n}} \exp\{\sum_{i=1}^k \theta_i T_i(\mathbf{y}^*)\}}, \quad \mathbf{y} \in \{0,1\}^{n\times n}. \tag{5}$$

For statistical inference, ERGMs are limited by computational and conceptual constraints, particularly with respect to out-of-sample inference as in the next scenario.

3.2 Modeling Friendships in a High School

Consider now a social network of friendships in a high school with N students, of which only $n < N$ are observed. We wish to use $\mathbf{Y}_n$ for sampled students to infer the structure of friendship patterns among all N students. Doing so requires an understanding of how sampled students are related to unobserved students in the population, raising questions about how the observed data $\mathbf{Y}_n$ represents the unobserved population $\mathbf{Y}_N$ and in turn how inferences about the sample relate to inferences about the population network. It is well known that this common statistical inference task poses conceptual difficulty for ERGMs [44], while other models, such as graphon models and the p_1 model, implicitly assume a "selection sampling" mechanism that is unrealistic for most practical applications; see Section 5.1 for an illustrative example of why this is the case. These observations raise doubts about the applicability and interpretability of statistical theory developed for certain network models, namely graphons, stochastic blockmodels and ERGMs, along with additional concerns about the practical implications of using these network models in modern applications. The modeling paradigm from Section 2 seeks to expand the scope of statistical network analysis beyond its current conceptual and theoretical limitations by accounting for both statistical variation and sampling in network analysis. Before exploring how specific model classes fit in this framework, we first review a few common network sampling mechanisms.

4. NETWORK SAMPLING

The following sampling descriptions represent specific instances of the inferential component within the framework of Section 2.

4.1 Selection Sampling

In the scenario of Section 3.2, suppose we model the population network $\mathbf{Y}_N$ as in (4). Given this description, how should we model the observed network $\mathbf{Y}_n$, $n < N$, so that the derived inferences for the population parameters ρ and β attain their intended meaning in terms of reciprocity and differential attractiveness? Is it legitimate to estimate ρ and β by first fitting (4) to $\mathbf{Y}_n$ and then using those estimates for the population $\mathbf{Y}_N$?

For concreteness, suppose the N students are labeled $1, \ldots, N$ and that $\mathbf{Y}_n$ is obtained by *selection sampling*: that is, $\mathbf{Y}_n$ is obtained from $\mathbf{Y}_N = (Y_{ij})_{1\le i,j\le N}$ by selecting the students labeled $1, \ldots, n$. If we assume $\mathbf{Y}_N$ is modeled by (4) with parameters $\rho, \theta, \alpha, \beta$, then the subsampled array $\mathbf{Y}_N|_{[n]} = (Y_{ij})_{1\le i,j\le n}$ is also distributed as in (4) with the same parameters [29]. Because the parameters for the sampled network $\mathbf{Y}_n$ preserve their interpretation under selection sampling, it is logically permissible to estimate the parameters governing $\mathbf{Y}_N$ by using the same estimated parameters as those obtained by fitting the p_1 model to the observed network $\mathbf{Y}_n$.

In the above modeling paradigm, (M1) and (M2) justify the above inferential protocol provided that:

(M1) the p_1 model adequately describes the variation in $\mathbf{Y}_N$ and

(M2) selection sampling adequately describes the relationship between $\mathbf{Y}_n$ and $\mathbf{Y}_N$.

More generally, let $(P_n)_{1\le n\le N}$ be a family of probability distributions with each P_n defined on $\mathcal{N}_n = \{0,1\}^{n\times n}$. We call $(P_n)_{1\le n\le N}$ *consistent under selection* if $\mathbf{Y}_n|_{[m]} \sim P_m$ for all $1 \le m \le n \le N$, that is, the random array obtained by first taking $\mathbf{Y}_n \sim P_n$ and then restricting to its first $m \le n$ rows and columns is distributed as P_m. Though a widely studied property of theoretical interest, for example, [44], consistency under selection has limited practical implications, as selection sampling is often not an accurate description of the relationship between data and population.

4.2 Generic Sampling

To describe more general sampling relationships between data $\mathbf{Y}_n$ and population $\mathbf{Y}_N$, we write $Y_n = \Sigma_{n,N}\mathbf{Y}_N$ to denote that $\mathbf{Y}_n$ was obtained from $\mathbf{Y}_N$ by sampling according to some (possibly random) sampling operation $\Sigma_{n,N}$. To be precise, $\Sigma_{n,N}$ is a randomly chosen function from the population space $\mathcal{N}_N$ to the sample space $\mathcal{N}_n$, where in general we write $\mathcal{N}_n$ for the space of networks of size n. (In typical settings, $\mathcal{N}_n = \{0,1\}^{n\times n}$ corresponds to the space of undirected graphs, but our framework accommodates network data of a more general form, as discussed above.) Therefore, when we refer to a *random sampling operation* $\Sigma_{m,n}$ from $\mathcal{N}_n$ to $\mathcal{N}_m$, we mean a randomly chosen map $\Sigma(\omega) : \mathcal{N}_n \to \mathcal{N}_m$,

TABLE 1
Database of movies and actors. Each row contains the set of actors in the corresponding movie. Based on [15], *Table* 3.2

	Starring cast
Rocky (1976)	Sylvester Stallone, Bert Young, Carl Weathers, ...
Rounders (1998)	Matt Damon, Ed Norton, John Turturro, ...
Groundhog Day (1993)	Bill Murray, Andie McDowell, Chris Elliott, ...
A Bronx Tale (1993)	Robert DeNiro, Chazz Palminteri, Joe Pesci, ...
Over the Top (1987)	Sylvester Stallone, Robert Loggia, ...
The Room (2003)	Tommy Wiseau, Greg Sestero, ...
⋮	⋮

where the underlying probability space $(\Omega, \mathcal{F}, P)$ is left implicit.[1] We reserve the notation $\mathbf{S}_{m,n}$, $m \le n$, to denote the operation of selection sampling a network of size m from one of size n as in Section 4.1. For example, for a network represented by a binary array $\mathbf{y} = (y_{ij})_{1\le i,j\le n}$, we write $\mathbf{S}_{m,n}\,\mathbf{y} := \mathbf{y}|_{[m]} = (y_{ij})_{1\le i,j\le m}$.

4.3 Edge Sampling

Suppose we sample calls from a telephone database in which each entry uniquely identifies caller and recipient of a different phone call. A sample of n calls (i.e., edges) results in a sequence of caller-receiver pairs $\{(C_i, R_i)\}_{1\le i\le n}$, which in a standard network representation is regarded as a graph with vertex set $\{C_1, R_1, \ldots, C_n, R_n\}$ and a directed edge from C_i to R_i for each $i = 1, \ldots, n$. It is clear that the structure induced by the observed interactions $\{(C_i, R_i)\}_{1\le i\le n}$ is not the result of selection sampling applied to the individuals (i.e., vertices) in the call log. The units are instead the phone calls (i.e., edges) and the sample size is the number of sampled edges. This example motivates the development of edge exchangeable models discussed in Section 9.

4.4 Hyperedge Sampling

Consider the Internet Movie Database (IMDB), whose entries correspond to different movies, their cast of actors, etc., as illustrated in Table 1. Sampling rows in this table corresponds to sampling different movies, which results in a sequence of movie casts $M_1, \ldots, M_n$ with each M_i recording the set of actors who play a role in the ith movie sampled, that is, $M_i = (M_{i,1}, \ldots, M_{i,R_i})$ where R_i is the number of roles in movie i. We observe similar structure when sampling academic articles from a research repository, such as arXiv. Article i is identified by its set of authors $A_i = (A_{i,1}, \ldots, A_{i,n_i})$ where n_i is the number of coauthors for article i. Authors are ordered according to convention of the scientific field in which the article is published. See Table 1 for an example database of movies and actors, where each row contains the set of actors in the corresponding movie; this table is based on [15], Table 3.2.

4.5 Path Sampling

One guiding question that motivated early interest in network science was to determine the structure of the Internet network by sampling the paths traveled by messages sent between servers. Traceroute [1] is a sampling method used for this purpose. Roughly, given a source s and target t, identified by their Internet Protocol (IP) addresses, traceroute returns the path (i.e., "traces the route") taken in accessing t from s. An observed network structure is obtained by piecing together the paths sampled by a number of applications of traceroute sampling.

4.6 Relational Sampling

The preceding examples of edge, hyperedge and path sampling, along with other common sampling schemes such as snowball sampling, are special cases of *relational sampling*, whereby a network is constructed from a sample of certain relations among individuals in a population. As network datasets are inherently relational structures and many sampling schemes encountered in practice are observed in a way that depends on the network structure itself, relational sampling is often more appropriate than the more vertex-centric procedures of selection sampling or simple random vertex sampling. Many other sampling schemes arise in practice (e.g., respondent driven sampling [31], subgraph sampling [33]), but those listed here are sufficient for illustrating the basic features of the above modeling framework.

5. MODEL COHERENCE

The description of a model as a family of distributions $\{\mathcal{M}_n\}_{1\le n\le N}$ together with a sampling context $\{\Sigma_{m,n}\}_{N\ge n\ge m\ge 1}$ provides the bare essentials for statistical inference as follows. From network data of size $n < N$, suppose we obtain a point estimate $\hat{P}_n \in \mathcal{M}_n$ based on a stated criterion, for example, maximum likelihood. Given $\hat{P}_n$ and the relationship between $\mathbf{Y}_n$ and $\mathbf{Y}_N$ established by the assumed sampling mechanism $\Sigma_{n,N}$, derive the estimate $\hat{P}_N \subseteq \mathcal{M}_N$ for the population as the set of all distributions that are consistent with $\hat{P}_n$ under sampling by $\Sigma_{n,N}$. In particular, any distribution P on $\mathcal{N}_N$ induces a distribution on $\mathcal{N}_n$ by first taking $\mathbf{Y}_N \sim P$ and then putting $\mathbf{Y}_n = \Sigma_{n,N} Y_N$, that is, the network obtained by sampling according to $\Sigma_{n,N}$. With $\Sigma_{n,N} P$ denoting this induced distribution, the inferred estimate for the distribution of $\mathbf{Y}_N$ based on $\hat{P}_n$ should be

$$\hat{P}_N = \{P : \Sigma_{n,N} P = \hat{P}_n\}. \tag{6}$$

[1] We consider such measure-theoretic considerations outside the scope of this article, which is aimed at an expository level.

The protocol in (6) can be expanded in an analogous way to other inferential objects, such as confidence regions. (Note that $\hat{P}_N$ is a singleton only if the model $\mathcal{M}_N$ is identifiable from $\mathcal{M}_n$ under sampling from $\Sigma_{n,N}$.)

Though standard in classical statistics, the inference step in (6) can be complicated in many network problems and other complex data applications. Deriving sensible inferences from a model $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\Sigma_{m,n}\}_{1\le m\le n\le N})$ requires a coherence among the model components (M1) and (M2) so that inferences based on the model $\mathcal{M}_m$ for data of size $m \ge 1$ have a clear logical relationship to inferences for the model $\mathcal{M}_n$ describing unobserved parts of the population. These considerations give rise to a natural coherence condition:

(C) For all $1 \le m \le n \le N$, $\Sigma_{m,n}\mathcal{M}_n = \mathcal{M}_m$,

where $\Sigma_{m,n}\mathcal{M}_n := \{\Sigma_{m,n}P : P \in \mathcal{M}_n\}$ is the model that $\mathcal{M}_n$ induces on $\mathcal{N}_m$ by sampling according to $\Sigma_{m,n}$.

In words, (C) requires that the asserted model $\mathcal{M}_m$ for each finite sample size $m \ge 1$ coincides with the models induced by sampling according to the assumed sampling scheme, that is, $\Sigma_{m,n}\mathcal{M}_n$ for all $n \ge m \ge 1$. This coherence condition is designed to ensure that inferences obtained from (6) admit a clear interpretation within the context of the chosen model. Given an estimate $\hat{P}_n \in \mathcal{M}_n$ and a sampling description $\Sigma_{n,N}$, the inferred population model $\hat{P}_N$ ought to be a subset of $\mathcal{M}_N$, which is assumed to describe the variation in the population network. On the other hand, any distribution $P \in \mathcal{M}_N$ for the population ought to induce a distribution on $\mathcal{N}_n$ that is in the model $\mathcal{M}_n$, or else the explicit description $\mathcal{M}_n$ is incompatible with the induced description via sampling. Together these considerations give two conditions, $\Sigma_{n,N}\mathcal{M}_N \subseteq \mathcal{M}_n$ and $\mathcal{M}_n \subseteq \Sigma_{n,N}\mathcal{M}_N$, which are equivalent to (C). We call a model $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\Sigma_{m,n}\}_{1\le m\le n\le N})$ *coherent* if it satisfies (C).

For a simple example of a coherent model, let $\mathcal{N}_n$ be $\{0,1\}^{n\times n}$, define $P_n(\cdot;\theta)$ to be the Erdős–Rényi distribution with parameter $\theta \in [0,1]$ on $\mathcal{N}_n$ as in (1) and let $\mathcal{M}_n = \{P_n(\cdot;\theta) : \theta \in [0,1]\}$ be the set containing all such distributions parametrized by $[0,1]$. In the context of selection sampling, the model $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\mathbf{S}_{m,n}\}_{1\le m\le n\le N})$ is coherent because the family of Erdős–Rényi distributions is consistent under selection. More explicitly, because each edge in $\mathbf{Y}_n \sim P_n(\cdot;\theta)$ is present independently with probability θ, the edges in the restriction $\mathbf{S}_{m,n}\mathbf{Y}_n = \mathbf{Y}_n|_{[m]}$ to the first m labeled vertices are also independent with probability p, so that $\mathbf{S}_{m,n}\mathbf{Y}_n \sim P_m(\cdot;\theta)$. From this, it follows that $\mathcal{M}_m = \mathbf{S}_{m,n}\mathcal{M}_n$, establishing (C).

To see the significance of this property for inference, consider how it affects point estimation for the parameter θ governing $\mathbf{Y}_N$ modeled by the Erdős–Rényi distribution in (1). Given a partial observation $\mathbf{Y}_n$ obtained from $\mathbf{Y}_N$ by selection sampling, we obtain a point estimate $\hat{\theta}_n = n^{-2}\sum_{1\le i,j\le n}\mathbf{Y}_{ij} \in [0,1]$ and $\hat{P}_n = P_n(\cdot;\hat{\theta}_n)$ for the distribution of $\mathbf{Y}_n$. By the relationship between $\mathbf{Y}_n$ and $\mathbf{Y}_N$ via selection sampling, and coherence of the models $\mathcal{M}_n$ and $\mathcal{M}_N$, this point estimate for $\mathbf{Y}_n$ gives an inferred distribution $\hat{P}_N = P_N(\cdot;\hat{\theta}_n)$ using the inference rule in (6).

For an example of an incoherent model, let $\mathcal{N}_n$ be $\{0,1\}^{n\times n}$ and $P_n(\cdot;\theta/n)$ be the Erdős–Rényi distribution with $\theta \in [0,1]$. Here, each edge occurs in $\mathbf{Y}_n$ independently with probability θ/n, and let $\mathcal{M}'_n = \{P_n(\cdot;\theta/n) : \theta \in [0,1]\}$ be the set containing all such distributions. The model $(\{\mathcal{M}'_n\}_{1\le n\le N}, \{\mathbf{S}_{m,n}\}_{1\le m\le n})$ is incoherent. Indeed, each edge in $\mathbf{Y}_n \sim P_n(\cdot;\theta/n)$ is present independently with probability θ/n, so the edges in the restriction $\mathbf{S}_{m,n}\mathbf{Y}_n = \mathbf{Y}_n|_{[m]}$ to the first m labeled vertices are independent with probability θ/n; however, $\mathbf{Y}_m \sim P_m(\cdot;\theta/m)$ has edges present independently with probability θ/m. The sets of distributions under $\mathcal{M}_m$ and $\mathbf{S}_{m,n}\mathcal{M}_n$ differ, with the former consisting of all Erdős–Rényi distributions on $\mathcal{N}_m$ parameterized by $[0,1/m]$ and the latter containing only those distributions parameterized by $[0,1/n]$.

To see the impact on inference, consider point estimation for the parameter θ governing $\mathbf{Y}_N$ modeled by $P_N(\cdot;\theta/N) \in \mathcal{M}'_N$. Given observation $\mathbf{Y}_n$ related to $\mathbf{Y}_N$ by selection sampling, the MLE is $\hat{\theta}_n = 1 \wedge (n^{-1}\sum_{1\le i,j\le n} Y_{ij}) \in [0,1]$ and $\hat{P}_n = P_n(\cdot;\hat{\theta}_n/n)$ for the distribution of $\mathbf{Y}_n$. By selection sampling, using the inference rule in (6), $\hat{P}_N = P_N(\cdot;\hat{\theta}_n/n)$, which may not be an element of $\mathcal{M}'_N$ on account of incoherence.

5.1 Sampling Consistency

As in the two examples above, most of the network science literature focuses on consistency under selection. However, a more general notion of consistency is needed to account for the variety of sampling mechanisms introduced in Section 4. The reason for considering the mechanisms is that selection sampling cannot capture the true sampling mechanism in many modern network applications. To see why, consider sampling $\mathbf{Y}_n$ by choosing a relatively small number of the indices uniformly at random from $\mathbf{Y}_N = (Y_{ij})_{1\le i,j\le N}$. Assuming the diagonal is 0, there are $N(N-1) \approx N^2$ possible nonzero entries in $\mathbf{Y}_N$. If $\sum_{1\le i,j\le N} Y_{ij} \approx \epsilon N$ for some constant $\epsilon > 0$ that does not depend on N and $\mathbf{Y}_n$ is obtained from $\mathbf{Y}_N|_S$ for a subset $S \subseteq [N]$ of $n \ll N$ vertices sampled uniformly at random, then each entry of $\mathbf{Y}_n = (Y^*_{ij})_{1\le i,j\le n}$ satisfies

$$\Pr(Y^*_{ij} = 1) \approx \epsilon N/((N(N-1)) \approx \epsilon/N, \quad 1 \le i \ne j \le n.$$

Thus, the probability that $\mathbf{Y}_n$ is nonempty satisfies

$$\Pr\Big(\bigcup_{1\le i\ne j\le n}\{Y^*_{ij} = 1\}\Big) \le \sum_{1\le i\ne j\le n}\Pr(Y^*_{ij} = 1) \approx n^2\epsilon/N,$$

which is negligible provided that n is sufficiently smaller than N. Any nontrivial observation $\mathbf{y} = (y_{ij})_{1\le i,j\le n}$ under the above simple random sampling scheme is thus a low probability event, raising questions about model adequacy. With this in mind, we propose the following general definition of consistency under arbitrary network sampling.

DEFINITION 5.1 (Sampling consistency). Given probability distributions P_m on $\mathcal{N}_m$ and P_n on $\mathcal{N}_n$, with $n > m$, and a generic sampling operation $\Sigma_{m,n}$ from $\mathcal{N}_n$ to $\mathcal{N}_m$, we call P_m and P_n *consistent with respect to* $\Sigma_{m,n}$ if $P_m = \Sigma_{m,n} P_n$, where $\Sigma_{m,n} P_n$ is defined as the distribution of $\Sigma_{m,n}\mathbf{Y}_n$ for $\mathbf{Y}_n \sim P_n$. In other words, P_m and P_n are consistent provided that drawing $\mathbf{Y}_m$ according to P_m yields the same distribution as first drawing from P_n and then subsampling according to $\Sigma_{m,n}$.

5.2 Invariance Principles and Sampling Schemes

Our emphasis on sampling context and model coherence establishes a necessary logical condition for out-of-sample statistical inference, as shown in (6) and Definition 5.1. In more traditional applications, the required coherence is implicitly specified by an assumed invariance principle, such as the classical i.i.d. and exchangeability assumptions. When modeling networks, such invariance principles often suggest a natural sampling context within which a model class is coherent. Below we discuss several common classes of network models described in terms of a characteristic invariance principle, for example, vertex exchangeability, relative exchangeability and edge exchangeability. We discuss the practical implications of the natural sampling context for these models. See [15], Chapters 3–5, and [41] for some other recent work on the relationship between sampling and invariance.

6. STATISTICAL NETWORK MODELS AND INFERENTIAL TARGETS

While the remaining sections focus on exchangeable network models, it should be noted that exchangeability is not an end goal in and of itself. Neither the *descriptive component* (M1) nor the *inferential component* (M2) require exchangeability, and overall exchangeability holds no special status in the proposed network modeling framework. In some cases, exchangeability is debatable. In others, it is entirely irrelevant. Indeed, there is good reason to challenge assumptions of exchangeability when additional structure is present, as is most often the case in networks applications. Given the complexity of statistical network models, however, invariance considerations yield statistical models with a built-in model coherence with respect to a particular sampling protocol. Exchangeability is one such invariance principle which can prove useful in preliminary investigations. At minimum, these models provide a benchmark for developing more intricate models.

6.1 Stability and Validity of Within-Sample Tasks: Beyond Data Reduction

For within-sample inferential tasks, the full specification of a coherent model may not be necessary. However, the general concept of *coherence* is still important. Consider, for example, a network scientist interested in within-sample cluster analysis. The scientist may argue that a vertex-centric model is better because no out-of-sample inference is involved in a clustering task. We would disagree. Any approach to such a task should be setup in a manner consistent with the underlying data generating process. Dempsey et al. (2019), for example, study cluster analysis of ArXiv articles to understand ArXiv topic overlap (i.e., to understand the interdisciplinary level of various topics). The results, presented briefly in Section 10.1, show that models built upon the interaction as the statistical unit yield substantial improvement in the clustering task over a vertex-centric approach, even in the absence of sampling concerns. The conclusion we reach is simple and universal: the units of observation are the units of observation, regardless of the inferential task.

The inferential task enters when assessing model adequacy. Indeed, most goals of statistical inference can be mathematically instantiated as functionals $F : \mathcal{P}(\mathcal{S}) \mapsto \mathbb{R}^k$ from the space of probability distributions on the sample space to some real-valued vector space $\mathbb{R}^k$ for some $k \ge 0$ [37]. Posterior predictive checks [24] are a Bayesian example of assessing model adequacy. The classical "Box's loop" [6] is a general iterative procedure that contains sequences of model formulation and criticism. Once the model is deemed adequate with respect to the inferential task, the task can be performed.

Within-sample cluster analysis is one of the simplest uses of a statistical model as a summary of the data. In this regard, cluster analysis can be regarded as *data reduction* of the observed network $\mathbf{Y}_{[n]}$ to a set partition $B : [n] \times [n] \to \{0, 1\}$, where $B(i, j) = 1$ implies unit $i \in [n]$ and $j \in [n]$ are in the same cluster. Often, however, the analyst desires to go beyond data reduction, and wishes to assess validity of the observed clusters (i.e., model adequacy in the parlance of Box's loop). One way to assess cluster validity is via stability and predictability [47]. As defined by [47], "stability refers to the acceptable consistency of a data result relative to 'reasonable' perturbations of the data or model." Examples of reasonable perturbations include bootstrap, jackknife and cross-validation. Completion of such analyses requires well-defined statistical units. Cross-validation, for example, requires data-splitting into training and test which implicitly requires well-defined statistical units. Chen and Lei [9], for example, propose a network cross-validation technique by vertex sampling; Li et al. [36], on the other hand, consider

network cross-validation by edge sampling. For prediction to be possible, the statistical model must be a family of processes. Thus application of predictability and stability to assess cluster validity requires the concept of statistical unit and coherence. General coherence would imply the binary relation $B : [n] \times [n] \to \{0, 1\}$ can be naturally extended to a function with domain $B : \mathbb{N} \times \mathbb{N} \to \{0, 1\}$. One instantiation of coherence that accounts for clustering is relatively exchangeable models in which the distribution $\mathbf{Y}$ is invariant to permutations σ that preserve the binary relation [18].

6.2 Out-of-Sample Inference and the Necessity of Model Coherence

A *statistical model*, traditionally defined [10, 35, 38], is a family of probability distributions $\mathcal{M}$ on the sample space $\mathcal{S}$ of all maps from the set of *statistical units* $\mathcal{U}$ into the *response space* $\mathcal{R}$. Some other authors discuss statistical modeling from various perspectives [21, 26, 38], but none of these prior accounts directly addresses the specific challenges of network modeling, namely the effects of sampling on network data and its subsequent impact on inference. The emphasis on model coherence is an emphasis on processes over distributions, which removes the distinction between estimation and prediction. Many statistical network models can be made coherent with respect to a suitably chosen sampling protocol; this coherence is only relevant in applications, however, if the sampling mechanism is interpretable. Vertex, edge, hyperedge, path and relational sampling are all idealized but interpretable sampling protocols.

Indeed, implicit in statements such as "valid out-of-sample inference" is the notion of *in-sample* and *out-of-sample* statistical units. Thus, if the scientist's ultimate goal is valid out-of-sample inference, one must have well-defined units which brings one back to considerations of model coherence and interpretable sampling protocols. If edges are the sampling units, then the response is a function $y : \mathcal{U} \to \text{fin}(\mathcal{P})$ where $\text{fin}(\mathcal{P})$ denotes finite multisets of the population $\mathcal{P}$. Therefore, the statistical network model will be edge-centric because the edges are the unit. A vertex-centrix statistical model can be fit to the observed sample as a means of constructing summary statistics (i.e., descriptive analysis of the observed data); however, this does not mean it makes sense as a means for building statistical models for inferential tasks.

7. VERTEX EXCHANGEABLE MODELS

Here, we specialize to the conventional setup in which vertices are the units and network data is represented by a $\{0, 1\}$-valued array $\mathbf{Y}_n$. The class of *vertex exchangeable* models in this section is characterized by the property of assigning equal probability to any two graphs that are isomorphic up to relabeling of their vertices, which in turn implies that the observed network reflects a representative sample of vertices from the population. Within the framework of Section 2, the natural sampling context for vertex exchangeable models is given by selection sampling or any vertex sampling scheme independent of the network, for example, simple random vertex sampling.

7.1 Finite Exchangeable Random Graphs

A random graph $\mathbf{Y}_N$ in $\{0, 1\}^{N \times N}$ is called *vertex exchangeable* if

$$\mathbf{Y}_N^\sigma =_{\mathcal{D}} \mathbf{Y}_N \quad \text{for all permutations } \sigma : [N] \to [N], \tag{7}$$

where $\mathbf{Y}_N^\sigma := (Y_{\sigma(i)\sigma(j)})_{1 \le i,j \le N}$ and $=_{\mathcal{D}}$ denotes *equality in distribution*. To describe finite exchangeable models, let $\mathcal{U}_N$ be the set of *unlabeled graphs* with N vertices, whose elements correspond to the equivalence classes of arrays $\{0, 1\}^{N \times N}$ under relabeling. More precisely, for $\mathbf{y}, \mathbf{y}' \in \{0, 1\}^{N \times N}$, we write $\mathbf{y} \cong \mathbf{y}'$ if there exists a permutation $\sigma : [N] \to [N]$ such that $\mathbf{y}^\sigma = \mathbf{y}'$. In this way, $\mathcal{U}_N$ consists of the equivalence classes over all $\mathbf{y} \in \{0, 1\}^{N \times N}$ given by $\langle \mathbf{y} \rangle_\cong := \{\mathbf{y}' \in \{0, 1\}^{N \times N} : \mathbf{y}' \cong \mathbf{y}\}$. One can easily show, for example, [15], Theorem 6.1 and Exercise 6.3, that any vertex exchangeable distribution for $\mathbf{Y}_N$ corresponds to a unique distribution $\mathbf{p}$ on $\mathcal{U}_N$ which first takes $U \sim \mathbf{p}$ in $\mathcal{U}_N$ and then, given $U = u$, puts $\mathbf{Y}_N = \mathbf{y}$ for $\mathbf{y}$ chosen uniformly among all $\mathbf{y}' \in \{0, 1\}^{N \times N}$ for which $\langle \mathbf{y}' \rangle_\cong = u$.

Though networks observed in practice are always finite, in many applications the population size N is either unknown or very large, on the order of millions or billions, making it both conceptually and computationally challenging to apply the theory of finite exchangeable models to real network data. In such cases, it is prudent to specify a model for $\mathbf{Y}^* = (y^*_{ij})_{1 \le i,j \le n}$ that is robust to the unknown population size $N \ge n$ from which the observation may have been drawn. With this, coherence (in the selection sampling context) requires that $\mathbf{S}_{n,N}\mathbf{Y}_N =_{\mathcal{D}} \mathbf{S}_{n,N'}\mathbf{Y}_{N'}$ for all $n \le N \le N'$, and in particular $\mathbf{Y}_N =_{\mathcal{D}} \mathbf{S}_{N,N'}\mathbf{Y}_{N'}$ for all $N \le N'$. In this case, a coherent model describes a family of random population structures $(\mathbf{Y}_N)_{N \ge 1}$, with each being exchangeable in the sense of (7) and coherent with respect to the selection sampling operations $(\mathbf{S}_{N,N'})_{1 \le N \le N'}$. These considerations lead to the class of graphon models.

7.2 Graphon Models

For $N \ge 1$, we define a model on $\{0, 1\}^{N \times N}$ by specifying a function $\phi : [0, 1] \times [0, 1] \to [0, 1]$, which can be assumed symmetric ($\phi(u, v) = \phi(v, u)$) when modeling undirected graphs. From ϕ, we construct a random array $\mathbf{Y}_N$ in $\{0, 1\}^{N \times N}$ by first drawing $U_1, \ldots, U_N$ i.i.d. Uniform[0, 1] and then, given $U_1, \ldots, U_N$, assigning the values Y_{ij} conditionally independently with probabilities

$$\Pr(Y_{ij} = 1 \mid U_1, \ldots, U_N) = \phi(U_i, U_j), \tag{8}$$

for all pairs $1 \le i, j \le N$. Intuitively, the random variables $U_1, \dots, U_N$ associate to each $i = 1, \dots, N$ a latent random effect U_i, allowing us to express the distribution of $\mathbf{Y}_N = (Y_{ij})_{1 \le i,j \le N}$ in closed form by

$$\begin{aligned} &\Pr_N(\mathbf{Y}_N = \mathbf{y}; \phi) \\ &\quad = \int_{[0,1]^N} \prod_{1 \le i,j \le N} \phi(u_i, u_j)^{y_{ij}} \\ &\qquad \times \big(1 - \phi(u_i, u_j)\big)^{1 - y_{ij}} \, du_1 \cdots du_N, \end{aligned} \tag{9}$$

for $\mathbf{y} = (y_{ij})_{1 \le i,j \le N}$. By construction, the family $(Y_n)_{n \ge 1}$ generated according to (8) with the same ϕ is consistent under selection. Thus, letting $\Phi = \{\phi : [0,1] \times [0,1] \to [0,1]\}$ be the set of all such functions and taking $\mathcal{M}_n = \{P_n(\cdot; \phi) : \phi \in \Psi\}$ for all $n \ge 1$ for some subset of functions $\Psi \subseteq \Phi$ determines an exchangeable model that is coherent with respect to selection sampling in the sense of Section 5. Aldous [3], pp. 124–125, refers to these processes as ϕ-processes, but here we adopt the modern terminology and call $\mathbf{Y}_N$ a *graphon process directed by* ϕ.

7.3 Statistical Implications of Vertex Exchangeability

The Aldous–Hoover theorem [2, 30] associates every infinite exchangeable random array $\mathbf{Y}$ to a probability measure φ on Φ such that $\mathbf{Y}$ can be constructed as a graphon process directed by ϕ chosen randomly according to φ. With this interpretation, the distribution of every exchangeable family of compatible arrays $(\mathbf{Y}_n)_{n \ge 1}$ is determined by a measure φ on the space Φ so that

$$\begin{aligned} &\Pr_n(\mathbf{Y}_n = \mathbf{y}; \varphi) \\ &\quad = \int_\Phi \Pr_n(\mathbf{Y}_n = \mathbf{y}; \phi) \varphi(d\phi), \quad \mathbf{y} \in \{0,1\}^{n \times n}, \end{aligned} \tag{10}$$

for $\Pr_n(\cdot; \phi)$ as defined in (9).

Equation (10) articulates the basic structure and limitations of vertex exchangeable models. Notice first that the Erdős–Rényi model with parameter $p \in (0,1)$ can be represented in this setting by the constant function $\phi(-,-) \equiv p$. It is well understood that the simple structure of this model is unable to replicate the heterogeneous features of many real-world networks, and consulting (10) suggests why graphon models may not offer much improvement over Erdős–Rényi in practical applications. The Aldous–Hoover theorem has two immediate practical implications for the use of graphon models in network analysis.

7.4 Edge Density

Define the *edge density* of a random array $\mathbf{Y} = (Y_{ij})_{i,j \ge 1}$ as the proportion

$$\epsilon(\mathbf{Y}) := \lim_{n \to \infty} \frac{1}{n^2} \sum_{1 \le i,j \le n} \mathbf{1}(Y_{ij} = 1). \tag{11}$$

By the Aldous–Hoover theorem, any exchangeable $\mathbf{Y}$ behaves as a graphon process for some randomly chosen ϕ. Given ϕ, the limit $\epsilon(\mathbf{Y})$ exists and is deterministic, allowing us to write

$$\epsilon(\phi) := \int_{[0,1]^2} \phi(u, v) \, du \, dv.$$

In many applications, the observed relational array $\mathbf{Y}_n$ for a sample $[n] \subseteq \mathbb{N}$ has empirical density $n^{-2} \sum_{1 \le i,j \le n} \times \mathbf{1}(Y_{ij} = 1)$ that is often judged to be "small" relative to n and thus leads to the assumption that $\mathbf{Y}$ is *sparse*, that is, that $\epsilon(\mathbf{Y}) = 0$ with probability 1. Under the graphon process construction, however, the limiting density $\epsilon(\phi)$ equals 0 only if $\phi(u, v) \equiv 0$ for almost all $u, v \in [0,1]$, and any $\mathbf{Y}$ generated from such a ϕ must have $Y_{ij} = 0$ for all $i \ne j$ with probability 1. This observation yields an important practical implication of the Aldous–Hoover theorem for vertex exchangeable random graphs:

A vertex exchangeable network model for a countable population is dense or empty with probability 1.

As discussed elsewhere, for example, [8, 15, 16, 42], this observation all but disqualifies graphons as a viable class of models in many modern contexts. This observation alone has inspired much recent work on new model classes, in particular the models discussed in Sections 8.1 and 9 below.

7.5 Representative Sample

Apart from the above empirical challenge, vertex exchangeability presents a conceptual issue for statistical applications. When the observed network $\mathbf{Y}_n$ is sampled from a larger population network of unknown size, vertex exchangeability implies that the observed vertices comprise a representative sample of *all* vertices. This assumption is necessarily violated in sparse networks, for which a representative sample of vertices would produce a network that is empty (i.e., has no edges) with high probability. We comment here, however, that while the focus of network modeling has been on devising new models which are able to replicate sparsity, this narrow focus may be misplaced. Vertex exchangeable, that is, graphon, models are a poor fit to real-world networks because they produce dense networks with homogeneous structure, but this defect of graphon models can be diagnosed prior to comparing the empirical properties of graphons to those of real-world networks. By simply noticing that the natural sampling context of graphon models, which takes the observed vertices to be representative of the population of all vertices, is incompatible with the way in which most network data is observed, the practical limitations of graphon models are immediately clear.

8. ALTERNATIVES TO VERTEX EXCHANGEABILITY

The viability of graphons suffers from three main challenges: (i) vertex exchangeability assumes homogeneity

of observed and unobserved vertices, (ii) the generality of the graphon setup raises identifiability and estimation issues and (iii) realizations from graphon models do not replicate empirical behaviors observed in many real world networks. These challenges have led to some alternative approaches to network modeling that aim to address heterogeneous network properties in a way that is statistically tractable. We review these below.

8.1 Sparse Graphon Models

An early approach to the above issues appeared in work by Bickel and Chen [4], who noted that the homogeneity of the ϕ-process is such that the number of edges grows on the order of n^2 as $n \to \infty$. Bickel and Chen subsequently propose to take $\mathbf{Y}_n$ as a ϕ_n-process with $\phi_n = \rho_n^{-1}\phi$ for some sequence $\rho_n \to \infty$ such that

$$\rho_n^{-1} \int_0^1 \int_0^1 \phi(u, v)\, du\, dv \to 0.$$

But while a sequence of graphs $(\mathbf{Y}_n)_{n\geq 1}$ distributed according to this model is sparse with probability 1, Bickel and Chen's formulation still maintains all of the same homogeneity properties that make graphon models unsuitable to most applications. Moreover, the class of models for $(\mathbf{Y}_n)_{n\geq 1}$ has not been shown to be coherent, in the sense of Section 5, with respect to any natural sampling scheme, raising the question of how inferences from this model are to be interpreted or how asymptotics proven in the proposed regime are to be used.

8.2 Graphex Models

A recent alternative approach [8] represents network data as a point process in the upper half-plane $[0, \infty) \times [0, \infty)$. For a motivating example, consider a network of Facebook friendships, which can be represented as a pair $(\mathbf{x}, \mathbf{y})$ where the entries of $\mathbf{x} = (x_i)_{i\geq 1}$ record the time at which user i first joined Facebook and $\mathbf{y} = (y_{ij})_{i,j\geq 1}$ is a binary array with $y_{ij} = 1$ if user i and j are friends and $y_{ij} = 0$ otherwise. This network can alternatively be represented by a point process $\mathbf{X} \subseteq [0, \infty) \times [0, \infty)$ with $(x_i, x_j) \in \mathbf{X}$ if and only if $y_{ij} = 1$. For $t > 0$, a finite random graph can be constructed by restriction of the point process $\mathbf{X}$ to $[0, t) \times [0, t)$; that is, $\mathbf{X}_t = \mathbf{X} \cap [0, t)^2$ and the random graph $\mathbf{Y}_t$ is constructed by restricting the binary array $\mathbf{y}$ to users i for which $x_i < t$.

In introducing this model, Caron and Fox [8], Section 3.5, suggest that each x_i be interpreted as "the time at which a potential node enters the network and has the opportunity to link with other existing nodes," as in the above scenario. The above example provides some motivation for the class of random graph models constructed from exchangeable random measures, for which Veitch and Roy [45] leverage Kallenberg's theory of exchangeable point processes to prove a *graphex representation*. The technical details of this model class are beyond the scope of this discussion. We simply point out here that graphex models, though based on the theory of exchangeable point processes, are not exchangeable with respect to vertices, nor edges, nor any simple, intepretable relations. Instead, ''exchangeability" here refers to the fact that any observation of the point process $\mathbf{X}$ over a time window of length $t \geq 0$ is representative of the process over any other window of length t. Due to space constraints, we cannot provide further technical details of this model class here; see [5, 8, 45] and more recent references for further discussion, and also [15], Chapter 7, for a discussion of the sampling interpretation of this exchangeability condition on point processes.

8.3 Relative Exchangeability

Neither graphon nor sparse graphon models are able to account for distinguishing characteristics in heterogeneous populations. For example, in a friendship network of n high school students there are cliques, that is, groups of students who are mostly friends with one another, and other heterogeneous patterns based on class year and extracurricular interests. In this case, the sampled vertices are still representative but are no longer homogeneous.

The *stochastic blockmodel* (SBM) [28] was originally proposed to handle such heterogeneity by partitioning vertices into nonoverlapping communities (or blocks) $B := (B_1, B_2, \ldots)$ so that two vertices v, v', one in block B_i and the other in block B_j, are related in $\mathbf{Y}$ with block-dependent probability p_{ij}. Notice that according to this description, the SBM is not vertex exchangeable, as the distribution of $\mathbf{Y}$ is not invariant under permutation of vertices occupying different blocks of B. The SBM is, however, invariant with respect to permutations of the vertices that leave the partition B unchanged. In particular, vertices can be interchanged with other vertices within their same community, but not with vertices in different communities. This property of the SBM illustrates a special case of the more general class of *relatively exchangeable* network models (cf. [18]) by which the distributional invariance properties are determined by the symmetries of an underlying structure in the population.

Another example of a relatively exchangeable model is the Hoff–Raftery–Handcock family of *latent space models* (LSMs) [27], which instead of a community structure allows for relationships among individuals to depend on covariates and their closeness in an abstract latent "social space." Let $Z = \{z_i\}_{i=1}^n$ denote the set of latent locations for each student where $z_i \in \mathbb{R}^k$ (i.e., a vector in some low-dimensional Euclidean space). Moreover, assume we observe some vector-valued characteristics $x_{i,j}$ which may be pair-specific. Under conditional independence given Z and $\{x_{i,j}\}_{1\leq i,j\leq n}$, the presence of an edge between i and j

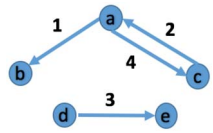


FIG. 1. *Network depiction of phone call sequence* (a,b), (c,a), (d,e), (a,c) *given in* (12).

FIG. 2. *Network depiction of phone call sequence* $X_1 = (a,b)$, $X_2 = (c,a)$, $X_3 = (d,e)$, $X_4 = (a,c)$ *along with its description under relabeling* X_2, X_3, X_4, X_1. *Any such reordering has equal probability under an exchangeable model.*

can be described by a conditional logistic regression, for example,

$$\text{logitpr}(y_{i,j} = 1 | z_i, z_j, x_{i,j}) = \alpha + \beta' x_{i,j} - |z_i - z_j|.$$

In a sense made precise in [18] relatively exchangeable models are a generalization of graphon models for which the edge probabilities depend not only on latent uniform random variables $U_1, U_2, \ldots$ as in (9) but also on the symmetries of some fixed underlying structure that determines the symmetries in the population. See [15], Chapter 8, for further discussion on the relationship between graphons and relatively exchangeable models.

9. EDGE EXCHANGEABLE MODELS

The models presented in Sections 7 and 8 take a vertex-centric perspective on network data, by which the natural units of observation either are the vertices or are closely related to the vertices (as in the case of graphex models). In many examples, as in those discussed in Sections 4.3–4.6, the edges or entities related to edges (such as hyperedges or paths) are better regarded as the units. This observation prompted the development of edge exchangeable models, for which we start with a simple motivating example. Within the framework of Section 2, the natural sampling context for edge exchangeable models is given by any edge sampling scheme as described in Section 4.3.

9.1 Sampling Phone Calls

Consider a network constructed by monitoring ingoing and outgoing calls at a telephone switchboard or by sampling calls at random out of a call log. With each call, we observe an ordered pair (s,r) corresponding to the *sender* s and *receiver* r of the call. Figure 1 depicts the structure of the sequence of calls

$$X_1 = (a,b), \quad X_2 = (c,a), \quad X_3 = (d,e), \quad X_4 = (a,c). \tag{12}$$

Notice the network representation labels the vertices a,b,c,d,e and edges $1,2,3,4$ so that the sequence X_1, X_2, X_3, X_4 can be exactly reconstructed. Networks constructed from email correspondence, scientific coauthorship, movie actor collaborations and trace-route sampling of Internet topology can all be formulated within a similar framework as this phone call example.

9.2 Edge-Centric Perspective

There is a tendency based on the representation in Figure 1 to regard the vertex labels a,b,c,d,e as arbitrary "names" which carry no additional meaning except to distinguish between vertices. While it is true that the vertex labels only identify individual vertices, it is crucial to note that the specification of vertex exchangeable models does not accurately reflect the manner in which such networks are observed. As Section 7 makes clear, vertex exchangeability brings with it implications well beyond the consideration of arbitrary vertex names. In the case of the sampled phone calls, the observed vertices are, in fact, part of the sampling process, and their identities (as determined by their phone call behavior) are also part of the data.

Assume the calls are sampled in such a way that the observations X_1, X_2, X_3, X_4 form an exchangeable sequence of ordered pairs. When regarded as the structure in Figure 2 with vertices identified, the data has the form of an exchangeable sequence. If modeled as an initial segment of an infinite exchangeable sequence $X_1, X_2, \ldots$, the available class of models is determined by de Finetti's theorem [19]. Even now, it is clear that delabeling the edges and viewing this as a vertex-labeled network imposes a perspective that does not reflect how the data is observed. It is reasonable, however, just as in Figure 2, to assume that the sequence of calls is sampled in an exchangeable manner, as illustrated in Figure 2. For the purpose of inference, then the sequences X_1, X_2, X_3, X_4 and X_1', X_2', X_3', X_4' determine the same interaction structure among vertices and thus convey the same information about the phone call database, even though individually the two observations may have different probabilities of occurrence. The net effect of this is the decision to disregard the vertex names in the network representation, as shown in Figure 3, leading to the edge exchangeable network models initiated and developed in [16].

9.3 Edge Exchangeable Models

The phone call data from Section 9.1 takes the form of a sequence $X_1, X_2, \ldots$ in the set $\mathcal{P} \times \mathcal{P}$ of ordered pairs in a population $\mathcal{P}$. Assume data comes in the form of a sequence $X_1, X_2, \ldots$ from $\mathcal{P} \times \mathcal{P}$. We may thus regard the data as a function $\mathbf{X} : \mathbb{N} \to \mathcal{P} \times \mathcal{P}$ in the usual way by the map $i \mapsto X_i$. For any bijection $\rho : \mathcal{P} \to \mathcal{P}$, we write $\rho\mathbf{X}$ to denote the composition of X with the function ρ

induces on $\mathcal{P}\times\mathcal{P}$ through $(a,b)\mapsto(\rho(a),\rho(b))$, so that $\rho\mathbf{X}:\mathbb{N}\to\mathcal{P}\times\mathcal{P}$ corresponds to $i\mapsto\rho(X_i)$. For any permutation $\sigma:\mathbb{N}\to\mathbb{N}$, we write $\mathbf{X}^\sigma:\mathbb{N}\to\mathcal{P}\times\mathcal{P}$ as the reordering of $\mathbf{X}$ according to σ, so that $\mathbf{X}^\sigma(i)=\mathbf{X}(\sigma(i))$. The *edge-labeled graph* induced by $\mathbf{X}:\mathbb{N}\to\mathcal{P}\times\mathcal{P}$ is defined formally as the equivalence class

$$\mathbf{y}_{\mathbf{X}}=\{\overset{\prime}{\mathbf{X}}:\mathbb{N}\to\mathcal{P}\times\mathcal{P}:\rho\overset{\prime}{\mathbf{X}}=\mathbf{X}\ \text{for some bijection } \rho:\mathcal{P}\to\mathcal{P}\}. \tag{13}$$

We write $\mathfrak{E}_{\mathbb{N}}$ to denote the space of edge-labeled graphs (with edges labeled in $\mathbb{N}$) defined as in (13), and $\mathbf{Y}$ to denote a random, edge-labeled graph. An illustration of the operation in (13) is shown in Figure 3.

Below we write $\mathbf{Y}^\sigma$ as the edge-labeled graph with edges relabeled according to a permutation $\sigma:\mathbb{N}\to\mathbb{N}$. Formally, this is defined by taking any $\mathbf{X}$ such that $\mathbf{Y}=\mathbf{y}_{\mathbf{X}}$ and putting $\mathbf{Y}=\mathbf{y}_{\mathbf{X}^\sigma}$, where $\mathbf{X}^\sigma$ is as defined above. A random edge-labeled graph is *edge exchangeable* if its distribution is invariant under this relabeling operation.

DEFINITION 9.1 (Edge exchangeability). A random edge-labeled graph $\mathbf{Y}$ is *edge exchangeable* if $\mathbf{Y}^\sigma=_{\mathcal{D}}\mathbf{Y}$ for all permutations $\sigma:\mathbb{N}\to\mathbb{N}$.

The Hollywood model is a special class of edge exchangeable models with statistical properties aimed at closing the gap between empirical aspects of network data and theoretical properties of statistical models. For (α,θ) satisfying either:

- $\alpha<0$ and $\theta=-k\alpha$ for some positive integer $k=1,2,\ldots$ or
- $0\le\alpha\le1$ and $\theta>-\alpha$,

the Hollywood model assigns probability to edge-labeled graphs with n edges

$$\begin{aligned}&\Pr(\mathbf{Y}_n=\mathbf{y};\alpha,\theta)\\&\quad=\alpha^{v(\mathbf{y})}\frac{(\theta/\alpha)^{\uparrow v(\mathbf{y})}}{\theta^{\uparrow(2n)}}\\&\qquad\times\prod_{k=2}^{\infty}\exp\{N_k(\mathbf{y})\log(1-\alpha)^{\uparrow(k-1)}\},\end{aligned} \tag{14}$$

where $v(\mathbf{y})$ be the number of nonisolated vertices in $\mathbf{y}$ and let $N_k(\mathbf{y})$ be the number of vertices in $\mathbf{y}$ with total degree (i.e., in-degree plus out-degree) k. The natural description of this model in terms of sampling movies from a movie database inspired the naming as the *Hollywood model* in [16].

The sparsity and power law properties of the Hollywood model are readily seen from the connection between (14) and the two-parameter Chinese restaurant process; see [12, 22] for a survey. In particular, for $\alpha\in(0,1)$ the degree distribution exhibits power law with exponent $\alpha+1$ and is sparse for $\alpha\in(1/2,1)$. When $0<\alpha<1$, $v(\mathbf{Y}_n)$ satisfies

$$E[v(\mathbf{Y}_n)]\sim\frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}n^\alpha \tag{15}$$

as $n\to\infty$, where '$a_n\sim b_n$ as $n\to\infty$' indicates that $\lim_{n\to\infty}a_n/b_n=1$ and $\Gamma(t)=\int_0^\infty u^{t-1}e^{-u}\,du$ is the gamma function; see [43], p. 69. From this, we deduce that the sequence $(\mathbf{Y}_n)_{n\ge1}$ obtained from the Hollywood model with parameter (α,θ) is sparse with probability 1 provided that $1/2<\alpha<1$, and the degree distribution has power law with exponent $\alpha+1$ when $0<\alpha<1$. See [16] for many more details about edge exchangeable models in general and the Hollywood model in particular.

9.4 Statistical Implications of Edge Exchangeability

In direct analogy to the Aldous–Hoover theorem and related results for vertex exchangeable models, Crane and Dempsey [16], Theorem 3.2, have proven a generic representation of edge exchangeable models as a mixture over what may be called *interaction propensity processes*. Define the $(\mathcal{P}\times\mathcal{P})$*-simplex* by

$$\mathcal{F}_{\mathcal{P}\times\mathcal{P}}=\Big\{(f_{(s,t)})_{s,t\in\mathcal{P}\times\mathcal{P}}:f_{(s,t)}\ge0\text{ and }\sum_{s,t\in\mathcal{P}}f_{(s,t)}=1\Big\}, \tag{16}$$

which corresponds to the set of all probability distributions on $\mathcal{P}\times\mathcal{P}$. Given any $f\in\mathcal{F}_{\mathcal{P}\times\mathcal{P}}$, we define ε_f as the probability distribution of a random edge-labeled graph $\mathbf{Y}\in\mathfrak{E}_{\mathbb{N}}$ determined by drawing $\mathbf{X}=(X_1,X_2,\ldots)$ i.i.d. from

$$P(X_i=(s,t)\mid f)=f_{(s,t)},\quad(s,t)\in\mathcal{P}\times\mathcal{P}, \tag{17}$$

and putting $\mathbf{Y}=\mathbf{y}_{\mathbf{X}}$ as defined in (13). The distribution of $\mathbf{Y}$ generated in this way is edge exchangeable by construction. We call ε_f the *interaction propensity process directed by* f. As an immediate observation, we note that the interaction propensity process allows for the occurrence of multiple edges, and in fact multiple edges between two vertices will occur with probability 1 in any large enough sample as long as $f_{(s,t)}>0$ for some $s,t\in\mathcal{P}$.

A second observation is that the vertices appear in $\mathbf{Y}\sim\varepsilon_f$ in size-biased order according to their overall frequency of occurrence in the network, thus demonstrating in a precise sense how the assumption of vertex exchangeability is incompatible with the manner in which the phone call data in Section 9.1 was observed. Whereas vertex exchangeability treats observed vertices as representative of the population of all vertices, edge exchangeability treats observed vertices as nonrepresentative (i.e., a size-biased sample) of all vertices. Many more details about edge exchangeable models are left to [16] and future work on this topic.

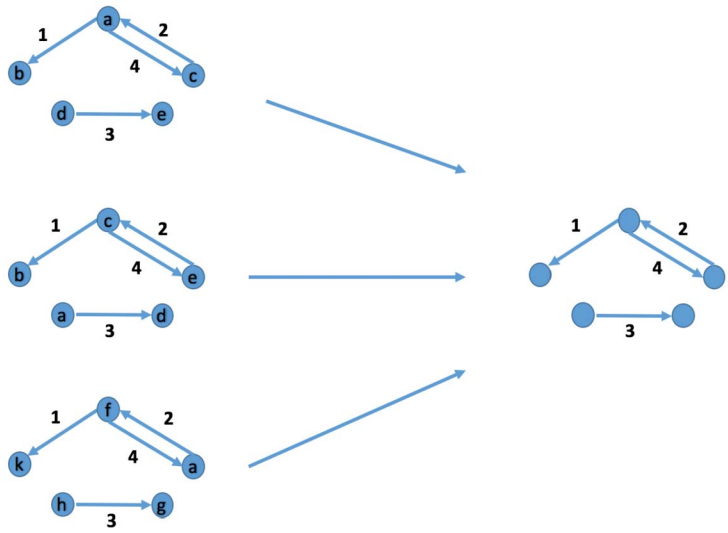


FIG. 3. *Phone call networks containing the same sufficient information, as shown by the edge-labeled graph on the right.*

10. BEYOND EDGE EXCHANGEABILITY

Many common network datasets admit a similar description to the motivating example of phone call sampling from Section 9.1. These examples suggest the refinement of edge exchangeability to relational exchangeability.

Like edge exchangeable models, relationally exchangeable models describe networks constructed by a representative sample of relations, for example, edges, hyperedges and paths, as described in Section 4. For example, in Section 4.5 we discussed how the Internet network can be sampled by piecing together paths traversed between randomly chosen source and target vertices. Following the same rationale as in the phone call example of Section 9.1, we realize that the different servers in the Internet do have distinct identities but that the identities of observed vertices cannot be severed from their participation in the observed paths. Under the assumption that the paths are representative of the population of all possible paths, we obtain a *path exchangeable network*, whose distribution assigns equal probability to path labeled graphs that are isomorphic up to relabeling of paths, and for which the analog to the representation of edge exchangeable models in Section 9.4 is a special case of the more general characterization of relationally exchangeable network models proven in [17].

As mentioned before, edge, hyperedge and path sampling are three special kinds of relational sampling, in which a network is obtained by sampling among some collection of relations. Because network datasets are primarily relational in nature, relational sampling is far more natural for most networks applications than the more commonly assumed vertex sampling. If the sampled relations are representative of a larger population of all relations, as in the presumed example of Section 9.1 or in sampling uniformly without replacement from the arXiv or SSRN, then the resulting network is called *relationally exchangeable*, as introduced and studied in [17]. Relationally exchangeable structures exhibit many analogous behaviors to the more specific edge exchangeable structures of Section 9. Because the concepts that arise for relational exchangeability are the same as those for edge exchangeability, but are more technical, we leave those details to [17].

10.1 Hierarchical Interaction Exchangeable Networks

We end this section by highlighting the impact of the edge-centric perspective on a real statistical network analysis. Here we consider a subset of articles posted to ArXiv. Relations such as scientific articles can be regarded as *hierarchical* interactions. Each article can be summarized by a list of scientific topics and a list of authors

$$X = (\bar{t}, \bar{a}) = \big((t_1, \ldots, t_{k_1}), (a_1, \ldots, a_{k_2})\big),$$

where $\bar{t}$ is the set of associated topics from a finite population of topics $\mathcal{P}_1$ and $\bar{a}$ are the associated authors from an infinite population of authors $\mathcal{P}_2$.

For simplicity, start by considering an article consisting of a single topic and single author $X = (t, a)$ and the following choice of interaction propensity process:

$$P\big(X = (t,a) \mid f^{(1)}, f^{(2)}\big) = f_t^{(1)} \times f_{a|t}^{(2)}, \tag{18}$$

where $f^{(1)}$ is a distribution on $\mathcal{P}_1$, and $(f_{\cdot|t}^{(2)})_{t\in\mathcal{P}_1}$ is a sequence of distributions on $\mathcal{P}_2$ indexed by $t \in \mathcal{P}_1$. The model is *hierarchical*, that is, the distribution of the author a depends on the topic t. The *hierarchical vertex components model* generalizes (18) to accomodate the more general structure of scientific articles; see [20] for details.

The statistical task is characterizing topic overlap within ArXiv. Two affinity matrices, denoted $A^{(e)}$ and $A^{(v)}$, are constructed with rows and columns indexed by the finite set of topics $t \in \mathcal{P}_1$. First, $A^{(e)}$ is constructed from the "edge-centric" perspective. For every pair of topics, t and t', an overlap score $SO(t,t')$ is calculated using the posterior distributions obtained under the hierarchical vertex components model and $A^{(e)}_{t,t'}$ is set equal to this score. The affinity matrix $A^{(v)}$ is constructed from the "vertex-centric" perspective. For every pair of topics, t and t', $A^{(v)}_{t,t'}$ is set equal to the number articles that contain both topics. A normalized spectral clustering algorithm [40] is applied to $A^{(e)}$ and $A^{(v)}$ with the number of clusters set to 6. Figure 4 reconstructs a heatmap that visualizes the spectral clustering analysis for topics that have been seen in the same article at least 100 times. Analysis of $A^{(e)}$ leads to clusters that are interpretable. Cluster 1, for example, includes cs.AI (Artificial Intelligence) and cs.IR (Information Retrieval); it is a group of topics that pertain to algorithmic approaches to artificial intelligence. Analysis of $A^{(v)}$, on the other hand, is unable to recover the meaningful groupings that the edge-centric approach produces. See [20] for additional discussion.

11. DYNAMIC NETWORK MODELS

We end by briefly mentioning some recent developments toward a theory for networks that are dynamic (i.e., whose connectivity changes over time). Consider, for example, the binary relational array of friendships from Section 3. From year to year, these friendships may change so that instead of observing just one array, we might observe the social relationships over the course of several years. In this case, the data is observed as a sequence $(\mathbf{y}(t))_{t\in\mathcal{T}}$ for some set of times $\mathcal{T}$. This section mostly considers the situation in which the vertex set is fixed while the edges are allowed to vary over time. The population process is therefore a collection of networks $\mathbf{Y} = (Y(t))_{t\in\mathcal{T}}$ indexed by times in $\mathcal{T}$, of which the entire population of N vertices or a sample of $n < N$ vertices may be observed.

11.1 Markov Property

Suppose that $\mathbf{Y}$ is a time homogeneous Markov chain, meaning that its transition behavior is governed by a family of transition probabilities $(P_t)_{t\in T}$ so that for any $\mathbf{y}, \mathbf{y}' \in \{0,1\}^{N\times N}$, the distribution of $\mathbf{Y}$ is given by

$$\begin{aligned}&\Pr\bigl(Y(t') = \mathbf{y}' \mid Y(t) = \mathbf{y}, \bigl(Y(u)\bigr)_{u\le t}\bigr)\\ &\quad := P_{t'-t}(\mathbf{y},\mathbf{y}'), \quad \mathbf{y},\mathbf{y}' \in \{0,1\}^{N\times N},\end{aligned} \tag{19}$$

for $t, t' \in \mathcal{T}$ and $t < t'$. In words, $P_{t'-t}(\mathbf{y},\mathbf{y}')$ is the probability that the network changes from $\mathbf{y}$ to $\mathbf{y}'$ in $t'-t$ units of time. While the Markov property is a common assumption for modeling dynamic phenomena in all manner statistical applications, there are good reasons to think that the Markov property is overly simplistic for many networks applications. Recall the descriptive component of the statistical model ought to account for the specific attributes of a given application. This is especially true for dynamic network models.

11.2 Temporal Exponential Random Graph Model (TERGM)

The most widely studied dynamic network model has so far been the *temporal exponential random graph model* (TERGM) [34], which is a natural temporal extension of the class of ERGMs discussed in Section 3. Let $\{0,1\}^{n\times n}$ be the state space for binary relational array and let Θ be some parameter space. In the TERGM, we define a joint sufficient statistic $T : \{0,1\}^n \times \{0,1\}^n \to \mathbb{R}^d$, where $d \ge 1$ is the length of the sufficient statistic vector $T = (T_1,\ldots,T_d)$. The TERGM defines the transition probabilities of a discrete time Markov chain $\mathbf{Y} = (Y(m))_{m\ge 0}$ by

$$\begin{aligned}&\Pr\bigl(Y(m+1) = \mathbf{y}' \mid Y(m) = \mathbf{y}; \theta\bigr)\\ &\quad \propto \exp\bigl\{\eta(\theta)\cdot T(\mathbf{y},\mathbf{y}')\bigr\}, \quad \mathbf{y},\mathbf{y}' \in \{0,1\}^{n\times n},\end{aligned} \tag{20}$$

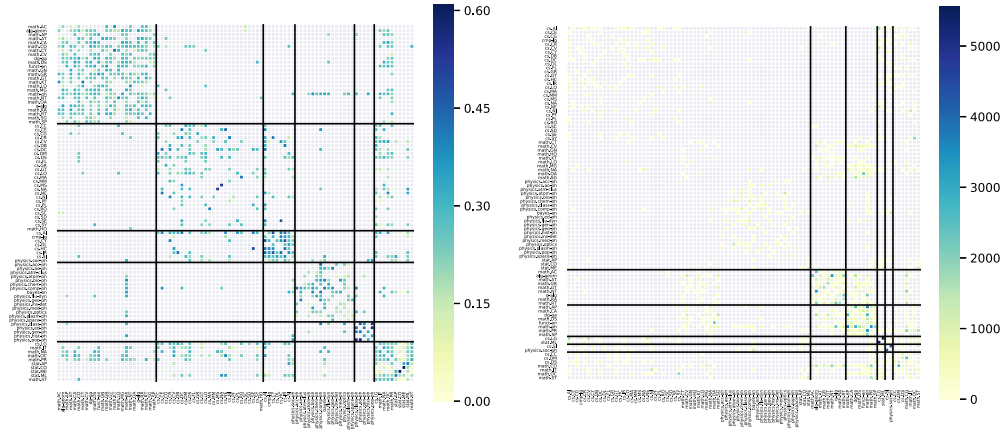
where $\eta(\theta)$ is the natural parameter for the exponential family. TERGMs incorporate Markovian dependence into the ERGM through the sufficient statistic T.

TERGM suffers from the the same practical issues as the ERGM. Kravitsky and Handcock (2014) presented the subclass of *separable TERGMs* (STERGMs) which describes network evolution with an intermediate step that *separates* the changes into a *formation model* and *dissolution model* for describing how edges are added and removed between times. TERGM may be appropriate when modeling the network dynamics for a fully observed population. In the case of modeling dynamics of a large network based on an observed sample, however, the model ought to account for the fact that observed dynamics are the result of a sampling process.

11.3 Projectivity and Sampling

Suppose a population process $\mathbf{Y} = (Y(t))_{t\ge 0}$ evolves on $\{0,1\}^{N\times N}$ for N assumed to be in the hundreds of millions, as in the Facebook network. While we are interested in learning the dynamics governing the entire network of size N, the population size is often too large to observe the while process all at once. We aim to infer the dynamics based on an observation from a sample of $n \ll N$ vertices, which need not preserve the Markov property. Here, we consider the case where sampling occurs by selection and consider circumstances under which the restricted processes $\mathbf{Y}_{[n]} = (Y_{[n]}(t))_{t\in T}$ are Markov for all $1 \le n \le N$, where we write $Y_{[n]}(t) := Y(t)|_{[n]}$.

In practice, this is an especially important property for modeling dynamic networks: since we must observe a



(a) Edge-centric analysis (b) Vertex-centric analysis

FIG. 4. *Heat maps of two-way entropy per article.*

sample $\mathbf{Y}_n$ of size $n \ll N$ from $\mathbf{Y}$, the dynamics associated to the observed data must be coherent with respect to the dynamics of $\mathbf{Y}$ and the manner in which $\mathbf{Y}_n$ was sampled, just as in Section 2. Otherwise the dynamics inferred from the observation $\mathbf{Y}_n$ would have no clear connection to the population process of interest.

Under the assumption that the Markov property of $\mathbf{Y}$ *projects* to each $\mathbf{Y}_{[n]}$ under selection, that is, $\mathbf{Y}_{[n]} = (Y_{[n]}(t))_{t\geq 0}$ is a Markov chain for every $1 \leq n \leq N$, it follows from a theorem of Burke and Rosenblatt [7] that the transition probabilities $P_t^{(n)}$ of $\mathbf{Y}_n$ can be expressed in terms of those of $\mathbf{Y}$ by

$$P^{(n)}(\mathbf{y}, \mathbf{y}') = P\big(\mathbf{y}^*, \{\mathbf{y}'' \in \{0,1\}^N : \mathbf{y}''|_{[n]} = \mathbf{y}'\}\big), \quad \mathbf{y}, \mathbf{y}' \in \{0,1\}^{n\times n}, \tag{21}$$

where $\mathbf{y}^* \in \{0,1\}^{N\times N}$ is any choice such that $\mathbf{y}^*|_{[n]} = \mathbf{y}$. Note that this condition is not always satisfied for the TERGM. We next discuss a class of models that do preserve the Markov property under selection sampling.

11.4 Rewiring Chains

For any $n = 1, 2, \ldots$, let $\mathcal{W}_n$ be the space of *rewiring maps*, which correspond to arrays taking values in $(\{0,1\} \times \{0,1\})^{n\times n}$. Each such array determines an operation on $\{0,1\}^{n\times n}$ in the following way. For any $W \in \mathcal{W}_n$, write the ij entry as $W(i,j) = (W_0(i,j), W_1(i,j))$ and define the image of $\mathbf{y} \in \{0,1\}^{n\times n}$ under W as the array $\mathbf{y}' = W(\mathbf{y}) = (y'_{ij})_{1\leq i,j\leq n}$ given by

$$\mathbf{y}'_{ij} = W_{y_{ij}}(i,j) = \begin{cases} W_1(i,j), & y_{ij} = 1, \\ W_0(i,j), & y_{ij} = 0. \end{cases} \tag{22}$$

In words, W acts on $\mathbf{y}$ by replacing each y_{ij} by the corresponding entry in W_1 if there is an edge ij in $\mathbf{y}$ and by W_0 if no edge ij in $\mathbf{y}$. For a concrete example:

$$\begin{matrix} \mathbf{y} \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix} \quad \begin{matrix} W \\ \begin{pmatrix} (\underline{\mathbf{0}}, 0) & (1, \underline{\mathbf{0}}) & (0, \underline{\mathbf{1}}) & (\underline{\mathbf{0}}, 0) \\ (1, \underline{\mathbf{0}}) & (\underline{\mathbf{0}}, 0) & (\underline{\mathbf{1}}, 0) & (\underline{\mathbf{1}}, 1) \\ (0, \underline{\mathbf{1}}) & (\underline{\mathbf{1}}, 0) & (\underline{\mathbf{0}}, 0) & (0, \underline{\mathbf{1}}) \\ (\underline{\mathbf{0}}, 0) & (\underline{\mathbf{1}}, 1) & (0, \underline{\mathbf{1}}) & (\underline{\mathbf{0}}, 0) \end{pmatrix} \end{matrix} \tag{23}$$

$$\mapsto \begin{matrix} W(\mathbf{y}) \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{matrix}.$$

We construct a Markov chain $\mathbf{Y} = (Y(m))_{m\geq 0}$ on $\{0,1\}^{N\times N}$ by first taking an initial state $Y(0) = \mathbf{y} \in \{0,1\}^{N\times N}$, a probability distribution ω on $\mathcal{W}_N$ and setting

$$\begin{aligned} Y(m+1) &= W_{m+1}\big(Y(m)\big) \\ &= (W_{m+1} \circ \cdots \circ W_1)(\mathbf{y}), \quad m \geq 0, \end{aligned} \tag{24}$$

for $W_1, W_2, \ldots$ drawn i.i.d. from ω. Following [11], we call the resulting process $\mathbf{Y}$ a *rewiring chain with directing measure* ω.

Notice that the transition behavior $\mathbf{Y}$ constructed in (24) is determined by the measure ω and that the action of each W on $\{0,1\}^{n\times n}$ as defined in (22) is such that the distribution of the restriction $Y(m+1)|_{[n]}$ given W_{m+1} and $Y(m)$ depends only on the restrictions $W_{m+1}|_{[n]}$ and $Y(m)|_{[n]}$. In particular, for any $W \in \mathcal{W}_N$, we define the restriction $W|_{[n]} \in \mathcal{W}_n$ in the usual way, $W|_{[n]} = (W(i,j))_{1\leq i,j\leq n}$. Then the restriction of $Y(m+1)$ to $\{0,1\}^{n\times n}$ is given by

$$Y(m+1)|_{[n]} = \big(W_{m+1}\big(Y(m)\big)\big)|_{[n]} = W_{m+1}|_{[n]}\big(Y(m)|_{[n]}\big).$$

It follows from this that all rewiring chains are projective by construction.

11.5 Exchangeable Rewiring Processes

The class of rewiring processes characterizes the behavior of projective Markov processes on $\{0,1\}^{\mathbb{N}\times\mathbb{N}}$ (countable population) whose transitions are *exchangeable*, in the sense that their transitions satisfy $P_t(\mathbf{y}^\sigma, \mathbf{y}'^\sigma) = P_t(\mathbf{y}, \mathbf{y}')$, for $\mathbf{y}, \mathbf{y}' \in \{0,1\}^N$ and $t \geq 0$ for all permutations $\sigma : [N] \to [N]$. For finite populations $\{0,1\}^{N\times N}$, an exchangeable discrete time chain $\mathbf{Y}$ can be constructed by taking ω to be an exchangeable probability measure on $\mathcal{W}_N$, meaning that $W \sim \omega$ satisfies $W^\sigma = (W(\sigma(i),\sigma(j)))_{1\leq i,j\leq N} =_{\mathcal{D}} W$ for all permutations $\sigma : [N] \to [N]$. If the population is taken to be countable and labeled in $\mathbb{N}$, then the class of rewiring chains characterizes the process.

THEOREM 11.1 (Crane [11, 13]). *Let* $\mathbf{Y} = (Y(m))_{m\geq 0}$ *be an exchangeable, projective Markov chain on* $\{0,1\}^{\mathbb{N}\times\mathbb{N}}$ *with initial state* $\mathbf{y} \in \{0,1\}^{\mathbb{N}\times\mathbb{N}}$. *Then there exists an exchangeable probability distribution* ω *on* $\mathcal{W}_{\mathbb{N}}$ *so that* $\mathbf{Y} =_{\mathcal{D}} \mathbf{Y}^* = (Y^*(m))_{m\geq 0}$ *generated by*

$$Y^*(m) = W_m\big(Y(m-1)\big) = (W_m \circ \cdots \circ W_1)(\mathbf{y}),$$

with $W_1, W_2, \ldots$ *are i.i.d. from* ω.

Theorem 11.1 limits the dynamics of exchangeable, projective Markov chains for countable graphs are limited to those with transitions that depend only locally on the current state of the process. For more discussion of graph-valued stochastic process models for dynamic networks, see [11, 13, 14] and [15], Chapter 11.

12. CONCLUDING REMARKS

We have presented a general framework within which to develop theory and methods for network analysis that accounts for the varied contexts in which such modern network data arise. Due to space constraints and the technical nature of this work, our presentation here was necessarily brief. There are, of course, many more aspects of statistical network analysis that have been treated elsewhere in the literature, for example, [15, 16], and which can be developed further in future work. The proposed statistical framework is intended to help the network science community move beyond the limitations of certain statistical models (e.g., graphons, ERGMs and SBMs) by providing guiding principles, that is, (M1), (M2) and (C), for building new models and methods to tackle the next generation of network science problems.

ACKNOWLEDGMENTS

Harry Crane is partially supported by NSF CAREER grant DMS-1554092.

REFERENCES

[1] ACHLIOPTAS, D., CLAUSET, A., KEMPE, D. and MOORE, C. (2005). On the bias of traceroute sampling or, power-law degree distributions in regular graphs. In *STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing* 694–703. ACM, New York. MR2181674 https://doi.org/10.1145/1060590.1060693

[2] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

[3] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—1983. Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. MR0883646 https://doi.org/10.1007/BFb0099421

[4] BICKEL, P. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[5] BORGS, C., CHAYES, J. T., COHN, H. and HOLDEN, N. (2017). Sparse exchangeable graphs and their limits via graphon processes. *J. Mach. Learn. Res.* **18** Paper No. 210, 71. MR3827098

[6] BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

[7] BURKE, C. J. and ROSENBLATT, M. (1958). A Markovian function of a Markov chain. *Ann. Math. Stat.* **29** 1112–1122. MR0101557 https://doi.org/10.1214/aoms/1177706444

[8] CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

[9] CHEN, K. and LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc.* **113** 241–251. MR3803461 https://doi.org/10.1080/01621459.2016.1246365

[10] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. CRC Press, London. MR0370837

[11] CRANE, H. (2015). Time-varying network models. *Bernoulli* **21** 1670–1696. MR3352057 https://doi.org/10.3150/14-BEJ617

[12] CRANE, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31** 1–19. MR3458585 https://doi.org/10.1214/15-STS529

[13] CRANE, H. (2017). Exchangeable graph-valued Feller processes. *Probab. Theory Related Fields* **168** 849–899. MR3663633 https://doi.org/10.1007/s00440-016-0726-0

[14] CRANE, H. (2018). Combinatorial Lévy processes. *Ann. Appl. Probab.* **28** 285–339. MR3770878 https://doi.org/10.1214/17-AAP1306

[15] CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis. Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

[16] CRANE, H. and DEMPSEY, W. (2018). Edge exchangeable models for interaction networks. *J. Amer. Statist. Assoc.* **113** 1311–1326. MR3862359 https://doi.org/10.1080/01621459.2017.1341413

[17] CRANE, H. and DEMPSEY, W. (2019). Relational exchangeability. *J. Appl. Probab.* **56** 192–208. MR3981153 https://doi.org/10.1017/jpr.2019.13

[18] CRANE, H. and TOWSNER, H. (2018). Relatively exchangeable structures. *J. Symbolic Logic* **83** 416–442. MR3835071 https://doi.org/10.1017/jsl.2017.61

[19] DE FINETTI, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036

[20] DEMPSEY, W., OSELIO, B. and HERO, A. (2019). Hierarchical network models for structured exchangeable interaction processes. Available at arXiv:1901.09982.

[21] DRTON, M. and SULLIVANT, S. (2007). Algebraic statistical models. *Statist. Sinica* **17** 1273–1297. MR2398596

[22] FENG, S. (2010). *The Poisson–Dirichlet Distribution and Related Topics*: *Models and Asymptotic Behaviors*. *Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2663265 https://doi.org/10.1007/978-3-642-11194-5

[23] FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

[24] GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

[25] HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082 https://doi.org/10.1214/08-AOAS221

[26] HELLAND, I. S. (2006). Extended statistical modeling under symmetry; the link toward quantum mechanics. *Ann. Statist.* **34** 42–77. MR2275234 https://doi.org/10.1214/009053605000000868

[27] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

[28] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[29] HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

[30] HOOVER, D. (1979). Relations on probability spaces and arrays of random variables. Institute for Advanced Studies. Preprint.

[31] KHABBAZIAN, M., HANLON, B., RUSSEK, Z. and ROHE, K. (2017). Novel sampling design for respondent-driven sampling. *Electron. J. Stat.* **11** 4769–4812. MR3729659 https://doi.org/10.1214/17-EJS1358

[32] KLUSOWSKI, J. and WU, Y. (2018). Counting motifs with graph sampling. *COLT* **75** 1966–2011.

[33] KLUSOWSKI, J. M. and WU, Y. (2020). Estimating the number of connected components in a graph via subgraph sampling. *Bernoulli* **26** 1635–1664. MR4091087 https://doi.org/10.3150/19-BEJ1147

[34] KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. MR3153932 https://doi.org/10.1111/rssb.12014

[35] LEHMANN, E. L. (1983). *Theory of Point Estimation*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0702834 https://doi.org/10.1007/978-1-4757-2769-2

[36] LI, T., LEVINA, E. and ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. MR4108931 https://doi.org/10.1093/biomet/asaa006

[37] MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320 https://doi.org/10.1214/aos/1035844977

[38] MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320 https://doi.org/10.1214/aos/1035844977

[39] MORENO, J. and JENNINGS, H. (1938). Statistics of social configurations. *Sociometry* **1** 342–374.

[40] NG, A. Y., JORDAN, M. I. and WEISS, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the* 14*th International Conference on Neural Information Processing Systems*: *Natural and Synthetic*, *NIPS'*01 849–856.

[41] ORBANZ, P. (2017). Subsampling large graphs and invariance in networks. Available at arXiv:1710.04217.

[42] ORBANZ, P. and ROY, D. (2014). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.

[43] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

[44] SHALIZI, C. R. and RINALDO, A. (2013). Consistency under sampling of exponential random graph models. *Ann. Statist.* **41** 508–535. MR3099112 https://doi.org/10.1214/12-AOS1044

[45] VEITCH, V. and ROY, D. (2015). The class of random graphs arising from exchangeable random measures. Available at arXiv:1512.03099.

[46] WASSERMAN, S. and PATTISON, P. (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and p. *Psychometrika* **61** 401–425. MR1424909 https://doi.org/10.1007/BF02294547

[47] YU, B. and KUMBIER, K. (2019). Three principles of data science: Predictability, computability. and stability (pcs). Submitted. Available at arXiv:1901.08152.