Title

Framework for converting mechanistic network models to probabilistic models

Journal

Journal of Complex Networks, 11(5)

ISSN

2051-1310

Authors

Goyal, Ravi

De Gruttola, Victor

Onnela, Jukka-Pekka

Publication Date

2023-09-05

DOI

10.1093/comnet/cnad034

Peer reviewed

Journal of Complex Networks (2023) **5**, Advance Access Publication on 16 October 2023
https://doi.org/10.1093/comnet/cnad034

Framework for converting mechanistic network models to probabilistic models

Ravi Goyal[†]
Division of Infectious Diseases and Global Public, Health, University of California San Diego, 9500 Gilman Drive, La Jolla, CA USA
[†]Corresponding author. Email: r1goyal@health.ucsd.edu

Victor De Gruttola
Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA USA

and

Jukka-Pekka Onnela
Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA USA

[Received on 7 December 2022; editorial decision on 16 August 2023; accepted on 25 August 2023]

There are two prominent paradigms for the modelling of networks: in the first, referred to as the mechanistic approach, one specifies a set of domain-specific mechanistic rules that are used to grow or evolve the network over time; in the second, referred to as the probabilistic approach, one describes a model that specifies the likelihood of observing a given network. Mechanistic models (models developed based on the mechanistic approach) are appealing because they capture scientific processes that are believed to be responsible for network generation; however, they do not easily lend themselves to the use of inferential techniques when compared with probabilistic models. We introduce a general framework for converting a mechanistic network model (MNM) to a probabilistic network model (PNM). The proposed framework makes it possible to identify the essential network properties and their joint probability distribution for some MNMs; doing so makes it possible to address questions such as whether two different mechanistic models generate networks with identical distributions of properties, or whether a network property, such as clustering, is over- or under-represented in the networks generated by the model of interest compared with a reference model. The proposed framework is intended to bridge some of the gap that currently exists between the formulation and representation of mechanistic and PNMs. We also highlight limitations of PNMs that need to be addressed in order to close this gap.

Keywords: networks; mechanistic models; probabilistic models.

1. Introduction

Conducting a randomized clinical trial (RCT) to evaluate the effectiveness of a prevention programme designed to mitigate the spread of disease may not be possible in many contexts due to logistical and financial complexities as well as the potentially long time frame required for a RCT [1–4]. Furthermore, it may be difficult to estimate effectiveness due to interference and spillover effects inherent in research on control of infectious diseases [5–9]. To address the challenges that arise in RCTs, there is increasing use of agent-based models (ABMs) for modelling effects of proposed prevention programmes; such efforts are facilitated by the advancement of high-speed computing resources [10–13]. A critical feature of ABMs is that their formulation facilitates simulation of interactions that can transmit disease among

the agents in the model; the collection of all such interactions within a population can be represented as a network. Investigators representing several different disciplines have developed specialized techniques for modelling these networks. Although there is potentially considerable synergy across the methods developed in different fields, limited tools currently exist to bridge them. In this article, we focus on bridging two of the primary techniques for generating simulated networks, mechanistic network models (MNMs) and probabilistic network models (PNMs).

MNMs generate a network by repeatedly applying a collection of stochastic microscopic rules. These rules can be simple, but nonetheless give rise to rich and complex network structure at the mesoscopic and macroscopic levels. There has been extensive research linking the presence or frequency of network structures to processes operating on a network, such as disease propagation [14–16]. For example, having a larger number triangles in a network can tend to decrease the size of an epidemic [17–19]. To statistically formulate and then address questions regarding whether mesoscopic- and macroscopic-level structures are more or less common in networks generated by mechanistic models than would be expected by chance typically requires models that specify a likelihood, denoted as $P_{\mathscr{G}}(G = g)$, of observing a given network g from a set of networks $\mathscr{G}$. We use the term PNM to describe such models; their importance arises from the way in which they can enable investigators to perform statistical inference. In this article, we propose a framework for specifying a PNM that is consistent with a mechanism model of interest in order to allow for statistical inference; we refer to this framework as mechanistic-to-probabilistic model conversion (MPMC).

In the next section, we provide an illustrative example of the connection between MNMs and PNMs as well as an example demonstrating that some commonly used PNMs are not suitable for this conversion. In Section 3, we discuss the theoretical reasons for such limitations. In addition, we discuss a fairly recently described PNM—referred to as the congruence class model (CCM)—that overcomes some of these limitations. Section 4 provides details of the proposed MPMC framework using CCMs and Section 5 provides two examples of this framework using a mechanistic model designed to provide insight into the HIV epidemic. In Section 6, we present an example that highlights the limitations in the use of PNMs (including CCMs) to model some mechanistic models. Section 7 discusses the proposed methods and suggests future research directions.

2. Background

A connection between PNMs and MNMs exists for specific sets of mechanistic rules. For example, let the generation of a network be governed by the mechanistic rule that individuals form edges with a fixed probability p and independent of all other edges. This generation process corresponds to the Erdős–Rényi–Gilbert model [20]; it also can be represented as an exponential random graph model (ERGM)—a common and flexible class of PNMs [21, 22]. Based on the authors' knowledge, this article provides the first framework for mechanistic to PNM conversion. Figure 1 illustrates our conceptualization of the connection between MNMs and PNMs; these two modelling paradigms are depicted as rectangles. The arrow connecting MNMs and PNMs (Arrow A) represents the subset of models wherein the association between the network generative mechanism(s) and corresponding probability distribution is known, such as for the ER model. For many models, the association between the mechanism(s) and probability distribution will not be obvious, and one must derive this connection through generating network realizations (data [circle in Fig. 1]) from the model: Arrows B and E in Fig. 1 represent starting from a MNM or PNM, respectively. From these generated data, it is possible to fit either a PNM (Arrow C) or MNM (Arrow D). This article focuses on converting a MNM to a PNM (Arrows B and C). Though ERGMs are quite

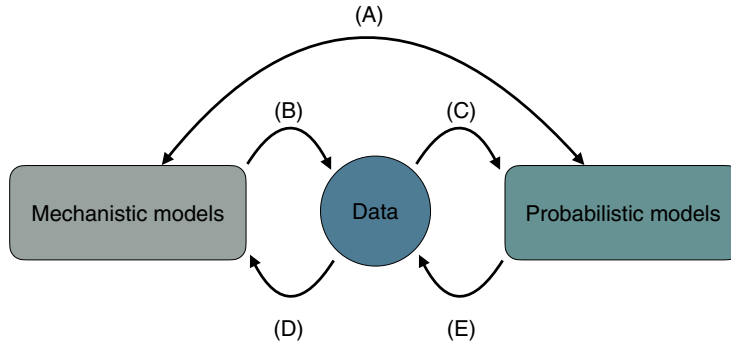

FIG. 1. Conceptual illustration of the conversion between mechanistic and PNMs: The mechanistic and probabilistic network modelling paradigms are depicted as rectangles. The arrow connecting MNMs and PNMs (Arrow A) represents the subset of models wherein the association between the network generative mechanism(s) and corresponding probability distribution is known. Arrows B and E represent generating network realizations (data [circle]) from a MNM or PNM, respectively. Arrows C and D representing fitting a PNM or MNM model, respectively, based on data. The MPMC framework is represented by Arrows B and C.

flexible, there are challenges to modelling MNMs using ERGMs (Arrow B then C in Fig. 1). The challenges are demonstrated through an investigation of a mechanistic model developed by Kretzschmar and Morris—hereafter referred to as the KM model [23, 24]—which played a significant role in identifying intervention priorities by highlighting the potential impact of concurrency on epidemic spread in sub-Saharan Africa [25]. In addition to its historic importance, the model continues to be the building block of more recently developed realistic models to study HIV [26]. This simple demonstration illustrates the need for a more flexible PNM than ERGMs.

2.1 *KM model*

Network evolution under the KM model is based on individual-level stochastic rules for partnership formation and dissolution. The population is fixed and the relationships among the population form and dissolve over time. At each time t, an individual can form new partnerships, dissolve existing partnerships or both. There are three key components governing the formation and dissolution of relationships: probability of pair formation (p_{f}), probability of pair separation (p_{s}) and a stochastic rule for partner mixing (ϕ), which can depend on the properties of the nodes. (Section 2.2 provides further details on the three key components of the KM model.) The evolution of a network under the KM model is outlined below:

1. Let g_t denote the network at time t.
2. Repeat the following T_t^1 times (T_t^1 is a KM model parameter):
   (a) Simulate a Bernoulli process where $X = 1$ with probability p_{f} and $X = 0$ otherwise.
   (b) If $X = 1$: (i) Draw two unconnected individuals at random: one male, i, and one female, j; (ii) with probability $\phi(i,j)$ add edge (i,j) to g; otherwise repeat (i) by redrawing two individuals at random.
3. Every connected node pair splits up with probability p_{s}.

The resulting network following these steps represents the network at time $t+1$, denoted as g_{t+1}.

To use the KM model to simulate an HIV epidemic, one must specify an initial network at time 0, denote this network as g_0. Once g_0 is specified, the steps outlined above can be used to generate networks at subsequent times. In the KM model, the network g_0 is generated by starting with an empty bipartite network with n_1 and n_2 nodes representing females and males, respectively, and then repeating the above steps a large number of times. This procedure is commonly referred to as a burn-in step. After completing this large number of iterations, the resulting network, g_0, is used at time 0. The burn-in step ensures that the simulation of the HIV epidemic starts at the stationary state of the network generation process. In Section 5, we provide examples of how the MPMC framework can be used to derive a PNM for the stationary state of the process. For the KM model, the stationary distribution, $P_{\mathscr{G}}(G = g)$, is unknown. However, once $P_{\mathscr{G}}(G = g)$ is known and given a method to sample from $P_{\mathscr{G}}(G = g)$, there is no need for the MCMC burn-in process going forward as one can sample a network g from the stationary state based on $P_{\mathscr{G}}(G = g)$. The MPMC framework can be used to derive a PNM for the stationary distribution of the KM model, i.e., derive $P_{\mathscr{G}}(G = g)$. However, the MPMC framework does require repeated simulation of a network from the stationary state using the burn-in process of the KM model. Therefore, the value of deriving a PNM identical to the KM model is reduction in future computational burden when applying or using the KM model. Note that in our article, we focus only on the generation of the networks and not on modelling the HIV epidemic on the networks.

2.2 *KM model and PNMs*

To illustrate the limitation of ERGMs to capture the KM model, we: (1) simulate k networks, $\{g_{M_1}, \ldots, g_{M_k}\}$, using a specification of the KM model; (2) sample k networks, $\{g_{S_1}, \ldots, g_{S_k}\}$, from an ERGM and (3) compare $\{g_{M_1}, \ldots, g_{M_k}\}$ to $\{g_{S_1}, \ldots, g_{S_k}\}$. The following provides additional details on each step.

Step 1: Simulate from KM

We investigate a simple specification of the KM model, pure random mixing, and use parameter values identical to those used by the authors of the KM model when it was first proposed [23]. In the setting of pure random-mixing, there exists no preference for nodes to form edges based on their covariates. The ϕ function for this setting is the following:

$$\phi(i,j) = \begin{cases} 1 \text{ if } k_i < d_m \text{ and } k_j < d_m \\ 0 \text{ else,} \end{cases} \tag{2.1}$$

where d_m is a KM model parameter and k_i and k_j are the current degrees of nodes i and j. The following parameters values were used in the original model: $n_1 = n_2 = 1,000$, $p_{\mathrm{f}} = 0.01$, $d_m = 10$ and $p_{\mathrm{s}} = 0.005$, $T_t^1 = (n_1 + n_2)/2 - |g_t|$, where $|g_t|$ is the number of edges in g_t.

Step 2: Simulate from ERGM

In Section 5, we provide evidence that the only network property necessary to represent the KM model for pure random mixing is the number of edges. Therefore, we consider ERGMs for which the number of edges is the only network statistic, reflecting our knowledge that other properties are not relevant. We investigate two ERGMs: (1) one that includes number of edges and (2) one that includes number of edges

and a constraint—implicit in the KM model—that the number of edges cannot exceed $(n_1 + n_2)/2$. The first ERGM has the following probability mass function (PMF):

$$P_{\mathscr{G}}(G = g|\omega) \propto \exp(\omega_1 \eta_1(g)), \tag{2.2}$$

where $\eta_1(g)$ is the number of edges in network g and ω_1 is the parameter associated with the number of edges. The second ERGM is similar, except in that the probability space with positive probability is restricted to networks with 1,000 or fewer edges. Therefore, the second ERGM has the following PMF:

$$P_{\mathscr{G}}(G = g|\omega) \propto \begin{cases} \exp(\omega_1 \eta_1(g)) \text{ if } |g| \leq 1,000 \\ 0 \text{ else,} \end{cases} \tag{2.3}$$

where $|g|$ is the number of edges in network g. For both ERGMs, we set ω_1 such that probability distribution is centred at the mean value for networks generated by the KM model.

Step 3: Comparison

We compare the cumulative density functions (CDFs) of the two collections of networks, $\{g_{M_1}, \ldots, g_{M_k}\}$ and $\{g_{S_1}, \ldots, g_{S_k}\}$, on network properties that consist of number of edges and number of individuals with degrees $\{0, 1, \ldots, 4\}$ (CDFs for degrees 5–10 are not shown here as few nodes had degrees in this range). The blue lines in Fig. 2 depict the CDFs of the network properties for the $k = 1,000$

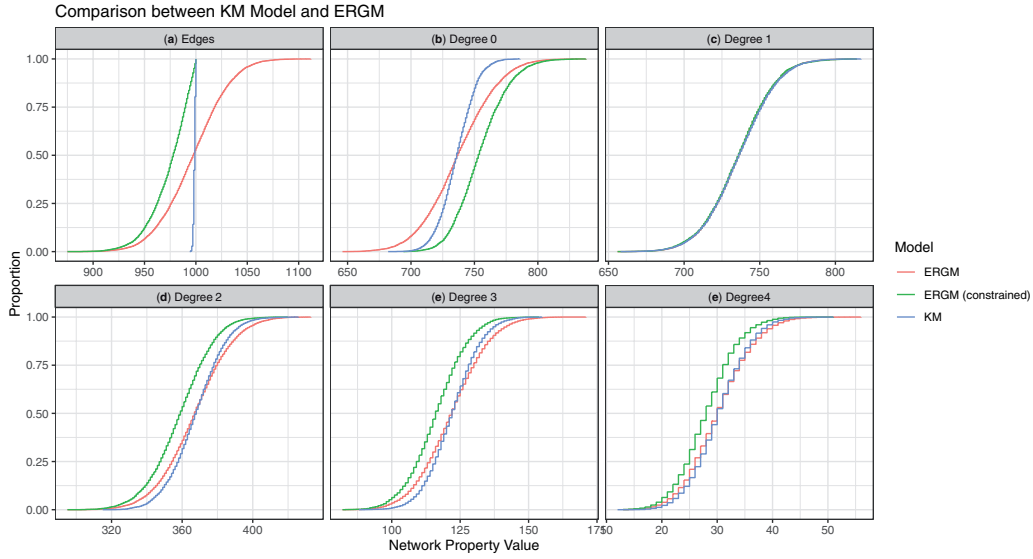

FIG. 2. Comparison between the KM model and ERGMs: A comparison of the number of edges and number of nodes of specified degree across the network collection for the KM model and ERGMs. Panel (a) depicts the CDF for the number of edges. Panels (b)–(f) depict the CDF for the number of nodes with degrees $\{0, 1, \ldots, 4\}$. The blue lines depict the CDFs for the KM model, and the red and green lines depict the CDFs for the ERGMs with and without the constraint on the number of edges, respectively. Because the CDFs generally do not match, the specified ERGMs are not able to capture the network structure generated by the KM model.

networks generated by the KM mechanistic model. The red and green lines depict the CDFs of the network properties for the k networks sampled from the ERGM with and without the constraint on the number of edges, respectively. The CDF associated with the KM model in blue is significantly steeper than is that for the ERGMs for the number of edges and for the number of nodes of degree 0; the CDF is only slightly steeper for the degrees greater than 0. The steeper CDFs for the KM model compared with those for the ERGM models indicate that the mechanistic model imposes additional constraints on the variability of the examined network properties compared with ERGMs. The two ERGMs used in this illustration are the only ERGMs that are possible assuming that the sole essential network property is the number of edges (evidence for this assumption is shown in Section 5). Therefore, for ERGMs to model the pure random mixing KM model, ERGMs would need to include more complex terms (and potentially a very large number of terms) to capture the KM model. The authors are not aware of any procedure to select these terms. Furthermore, inclusion of these complex terms would provide an incorrect interpretation of the generative process associated with the KM model for pure random mixing.

The PMF on the space of networks of size n for an ERGM is solely based on constraints, each of which specifies the expected frequency of a network structure of a random network sampled based on an ERGM. However, the KM model has an additional mechanism for creating edges. Specifically, the network generated at time step $t+1$ is based on the number of edges present at time step t. This mechanistic rule creates a dependency among all of the edges in a network, i.e., the presence of an edge in a network generated by the KM model depends on the presence or absence of all other edges. This dependency can be difficult to capture within the ERGM framework. In the next section, we provide details of ERGM theory to highlight this issue.

3. Previous work

To highlight some theoretical limitations of representing mechanistic models using ERGMs, we provide technical details for deriving the ERGM probability distribution. In Section 3.2, we present a recent network model that overcomes some, but not all, of these limitations; we return to this discussion in Section 6.

3.1 *Limitations of ERGMs*

In positing an ERGM, i.e., specifying $P_{\mathscr{G}}$, one proposes a dependence hypothesis that defines contingencies among the network edges, which are regarded as random variables; each potential edge, E_{ij}, has a corresponding random variable, denoted as X_{ij} [27]. This hypothesis can be codified through the specification of a dependence graph, denoted as $G_{\mathrm{D}} = (V_{\mathrm{D}}, E_{\mathrm{D}})$, on a population V. The nodes of G_{D} are tuples (i,j), where $i,j \in V$. An edge in G_{D} is represented as a pair of tuples, i.e., $\{(i,j),(k,m)\}$, where $i,j,k,m \in V$. Here, $\{(i,j),(k,m)\}$ is an edge in G_{D} if and only if edges (i,j) and (k,m) are conditionally dependent given information on all other potential edges, that is, the probability of the edge (i,j) existing in a network depends on the presence of edge (k,m). Let C denote the set of cliques (a subset of vertices such that every two distinct vertices are connected) in G_{D}; the cliques can be of any size. Let G_c be the graph formed by the collection of all edges denoted by the nodes of $c \in C$; Fig. 3 provides an illustration of a clique c and corresponding subgraph G_c.

The Hammersley–Clifford theorem states that $P_{\mathscr{G}}$ is a Gibbs distribution that can be factored over G_{D}, conditional on $P_{\mathscr{G}}$ being a positive distribution, i.e., $P_{\mathscr{G}}(G = g) > 0$, for all $g \in \mathscr{G}$ [28]. Therefore,

$$P_{\mathscr{G}}(G = g) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c), \tag{3.1}$$

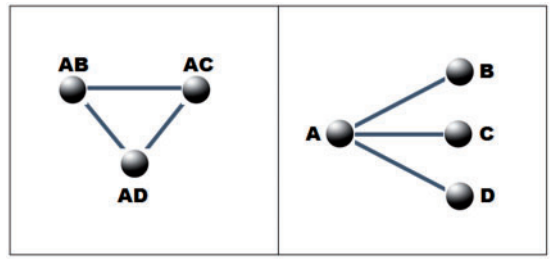

FIG. 3. Illustration of a clique: an illustration of a clique c in the dependence graph is shown in the left panel. The corresponding subgraph G_c is shown in the right panel.

where ψ_c is a function over sets of variables X_c associated with clique c in G_D and Z is a normalizing constant. As Equation (3.1) does not provide a unique distribution, additional constraints are necessary. A natural set of constraints assigns the probability of observing G_c for each $c \in C$. These constraints control the probability of observing a subgraph in which the occurrence of each edge depends on all of the other edges; the constraints are represented in Equation (3.2):

$$\sum_{g \in \mathcal{G}} I_{G_c \subseteq g} P_{\mathcal{G}}(G = g) = P_{\mathcal{G}}(I_{G_c \subseteq g}), \tag{3.2}$$

where I_{G_c} is the indicator function that G_c is a subgraph of g and $P_{\mathcal{G}}(I_{G_c \subseteq g})$ is the probability that needs to be specified.

As all subgraphs of G_c are associated with a clique in C, they too would be the subject of a constraint. Even with these constraints, $P_{\mathcal{G}}$ is not uniquely defined. In order to specify $P_{\mathcal{G}}$, ERGMs use the probability distribution that maximizes the Shannon entropy subject only to constraints represented in Equation (3.2); the maximum entropy principle is conceptually powerful and has numerous applications in science—particularly in physics [29]. The maximum entropy distribution best represents the current state of knowledge of a system, while assuming maximal ignorance about the distribution other than what is imposed by Equation (3.2) [30, 31]. This approach leads to the following distribution:

$$P_{\mathcal{G}}(G = g) = \frac{1}{Z} \prod_{c \in C} \exp(\omega_{G_c} I_{G_c \subseteq g}), \tag{3.3}$$

where ω_{G_c} is a parameter used to fix the mean probability of observing G_c, i.e., specify $P_{\mathcal{G}}(I_{G_c \subseteq g})$. Therefore,

$$\psi_c(X_c) = \exp(\omega_{G_c} I_{G_c \subseteq g}). \tag{3.4}$$

As the distribution specified in Equation (3.3) has a large number of parameters, $\{\omega\}$, one can simplify the model by imposing a homogeneity assumption that sets parameter values equal when they refer to the same type of subgraph, e.g., edge pairs and triangles. The resulting PMF presented below is the standard form for ERGMs:

$$P_{\mathcal{G}}(G = g|\omega) \propto \exp(\omega^T \eta(g)), \tag{3.5}$$

where ω is a (column) vector of model parameters associated with the specified network properties and $\eta(g)$ denotes the vector of counts for the network configuration associated with the cliques in G_D (also

referred to as sufficient network statistics for the ERGM), i.e., $\eta : \mathscr{G}_n \to \mathbb{R}^p$, where p is the length of the vector. As referred by Cimini *et al.* [31], ERGMs are examples of a canonical approach; that is, an approach in which networks are generated to have network features that match the observed network in expectation. This is in contrast to microcanonical approaches, which generate networks that exactly match observed network properties—for example, the configuration model [30, 32].

In developing the PMF for ERGMs, there are two critical requirements. The first is that the dependence graph, G_{D}, not be complete. A complete dependence graph results in $2^{\binom{n}{2}}$ cliques, which not only causes there to be a large number of parameters in Equations (3.3) and (3.5) but also creates identifiability issues; dense dependence graphs may also be problematic for a similar reason. The second requirement is that Equation (3.2) represents the only constraints on the system; that is, only the mean of the configuration counts is constrained. This precludes inclusion of information on the second or third moments on network configurations. For instance, Equation (3.2) allows neither specification of uncertainty in those counts (measurement error) nor variability around those counts (due to the stochastic nature of the mechanistic rules). The KM model violates these ERGM requirements.

3.2 *Congruent class model*

Because of their limitations, ERGMs cannot be used to represent the KM mechanistic model; this inability illustrates the need for greater flexibility in the modelling of network properties. To overcome some of these limitations, we propose an MCMC framework that uses the congruent class model (CCM) [33]. The class of models allows for greater flexibility in specifying the functional form of the probability distributions associated with network properties.

The CCM partitions the space of networks on n nodes, $\mathscr{G}$, such that all networks within a partition have the same values for the network properties of interest; these partitions are referred to as congruence classes. For example, one congruence class might correspond to all networks with 50 closed triads; another, to all networks with 51 closed triads and so on. Hence, a congruence class is defined as $c_x = \{g : \eta(g) = x, g \in \mathscr{G}_n\}$, where $\eta(g)$ denotes the value of the properties used to define the congruence classes for $g \in \mathscr{G}$. The number of networks in c_x is denoted as $|c_x|$. The probability distribution on $\mathscr{G}$ for the CCM is based on specifying $P_{\mathscr{C}}$, the PMF for the congruence classes defined by the essential network properties; $P_{\mathscr{C}}(x)$ is the total probability of all networks that are elements in c_x:

$$P_{\mathscr{C}}(x) = \sum_{g \in c_x} P_{\mathscr{G}}(g). \tag{3.6}$$

Because the congruence classes represent the partition of the space $\mathscr{G}$ based on essential network properties, two networks within a congruence class must have the same probabilities of being observed. Therefore, the probability distribution on $\mathscr{G}$ for the CCM is the following:

$$P_{\mathscr{G}}(G = g) \propto \left(\frac{1}{|c_{\eta(g)}|}\right) P_{\mathscr{C}}(c_{\eta(g)}). \tag{3.7}$$

For additional details on CCMs including a comparison with ERGMs, see [34, 35].

4. Framework

We denote the network generation rules of a MNM as γ. Though mechanistic models do not explicitly specify a PMF on a set of networks, they do so implicitly. Let $P_{\mathscr{G}}(G = g|\gamma)$ denote this implicit PMF,

where G is a random variable with support on $\mathscr{G}$ and $g \in \mathscr{G}$. Let $P_{\mathscr{G}}(G = g|\omega)$ denote a PMF for a PNM, where ω represents functions or parameters necessary to specify the PMF; this formulation allows for the PMF to be parametric, semi-parametric or non-parametric. We consider a collection of network properties to be *essential* for a model if the omission of any one property makes it no longer possible for the collection to characterize the model. The goal of MPMC is to uncover the essential network properties and their joint probability distribution, such that the probability of observing a network g is identical, whether the network is generated from the mechanistic model with rules γ or is sampled from a probabilistic model with parameter ω.

The general MPMC framework is an iterative algorithm; an outline of the conversion framework is as follows:

1. *Simulate the mechanistic model:* Generate a collection of networks, $\{g_{M_1}, \ldots, g_{M_k}\}$, based on simulating the mechanistic model k times.
2. *Propose essential network property candidates:* Based on subject matter knowledge, conceptual knowledge of the mechanisms, and previous iterations of the algorithm, propose a collection of network properties, defined by the function η, as the essential network properties of the mechanistic model.
3. *Estimate the joint probability distribution of essential network properties:* Estimate the joint probability distribution, $P_{\mathscr{C}}(s_1, ..., s_j)$, of the candidate essential network properties, defined by η, based on the observed simulated networks $\{g_{M_1}, \ldots, g_{M_k}\}$. In high dimensions, i.e., settings where a large number of network properties is being considered, density estimation is a non-trivial problem. However, given the generic nature of the problem, there exists a vast literature on methods for density estimation in this setting [36, 37].
4. *Sample networks:* Sample networks, $\{g_{S_1}, \ldots, g_{S_k}\}$, based on a CCM with the estimated joint probability distribution $P_{\mathscr{C}}(s_1, ..., s_j)$.
5. *Compare networks:* Statistically compare the probability distribution of the two collections of networks, $\{g_{M_1}, \ldots, g_{M_k}\}$ and $\{g_{S_1}, \ldots, g_{S_k}\}$, on a large set of network properties defined by η' not contained in the set defined by η.
6. *Iterate:* If statistical tests do not reject the hypothesis that the probability distributions on each of the network properties defined by η' differ between $\{g_{M_1}, \ldots, g_{M_k}\}$ and $\{g_{S_1}, \ldots, g_{S_k}\}$, then accept the properties defined by η as the essential network properties, such that their joint probability distribution $P_{\mathscr{C}}$ characterizes the network properties induced by the mechanistic model. Otherwise, repeat steps 2–6.

5. Application

In this section, we investigate the KM model described in Section 2. Specifically, we investigate two rules for partner mixing: serial monogamy and pure random mixing. In neither setting is it straightforward to understand what the mechanistic rules of the KM model imply about the properties of the induced networks. We use identical parameter values as the authors of the KM model when it was first proposed [23] and shown in Section 2.

5.1 *Pure random mixing*

To characterize the pure random mixing setting of the KM model, i.e., to identify the essential network properties of the mechanistic model along with their joint probability distribution, we follow the steps

of the MPMC framework outlined in Section 4. As described in Section 2, in the random mixing setting, there exists no preference for individuals to form relationships based on degree.

1. *Simulate the mechanistic model:* Let γ_r denote the microscopic rules associated with the pure random mixing setting for the KM model. We simulate $k = 10,000$ networks, $\{g_{M_1}, \ldots, g_{M_k}\}$, based on γ_r.
2. *Propose essential network property candidates:* Based on Fig. 2 it may appear that it would be necessary to model the degree distribution; however, we propose modelling only number of edges as the essential network property of the KM model. Let $X_E^{\gamma_r}$ represent the random variable for the number of edges in a network generated with γ_r.
3. *Estimate the joint probability distribution of essential network properties:* Let $P_E^{\gamma_r}$ denote the PMF for $X_E^{\gamma_r}$. From the blue line in panel (a) of Fig. 2, it appears that the distribution $P_E^{\gamma_r}$ does not follow any common distribution; therefore, we estimate $P_E^{\gamma_r}$, denoted as $\hat{P}_E^{\gamma_r}$, by letting $\hat{P}_E^{\gamma_r}(X_E^{\gamma_r} = x)$ equal the fraction of the k generated networks that have x edges, i.e., $\hat{P}_E^{\gamma_r}(X_E^{\gamma_r} = x) = \frac{1}{k}\sum_{i=1}^{k} I_{\eta(g_{M_i})=x}$.
4. *Sample networks:* We sample $10,000$ networks based on the following PMF:

$$P_{\mathscr{G}}(G = g|P_E^{\gamma_r}) \propto \left(\frac{1}{|c_{\eta(g)}|}\right) P_{\mathscr{C}_r}(c_{\eta(g)}), \tag{5.1}$$

   where $P_{\mathscr{C}_r}(c_{\eta(g)}) = \hat{P}_E^{\gamma_r}(X_E^{\gamma_r} = \eta(g))$ and $\eta(g)$ is the number of edges in g.
5. *Compare networks:* Figures 4 and 5 compare the networks generated from the KM model and those sampled from the CCM based on Equation (5.1) on a large set of network properties which consists of the number of edges, number of nodes of degree 0–4 (nodes of higher degree were extremely rare), betweenness centrality (max and mean across all nodes), degree correlation, eigenvalue centrality (max and mean across all nodes) and number of K-stars (1–3); detailed descriptions of the metrics are available in [38] and [30]. Based on the Kolmogorov–Smirnov test, one cannot reject the hypothesis that the network property distributions are identical (the p-values ranged from 0.23 to 1 across all of the network properties) [39].
6. *Iterate:* Based on the Kolmogorov–Smirnov tests, we conclude that the number of edges is the only essential network property and the probability distribution in Equation (5.1) characterizes the mechanistic random mixing KM model.

5.2 *Serial monogamy*

In the serial monogamy setting, individuals are restricted from having more than one partner at the same time. In the article by Kretzschmar and Morris [23], the ϕ function for this setting is the following:

$$\phi(x, y) = \begin{cases} 1 \text{ if } k_i = k_j = 0 \\ 0 \text{ else.} \end{cases} \tag{5.2}$$

For the remaining parameters, we use values that are identical to those used by the authors of the KM model when it was first proposed [23] (see Section 2).

As in the previous example, to characterize the serial monogamy setting of the KM model, i.e., identify the essential network properties of the mechanistic model along with their joint probability distribution, we follow the steps of the MPMC framework outlined in Section 4.

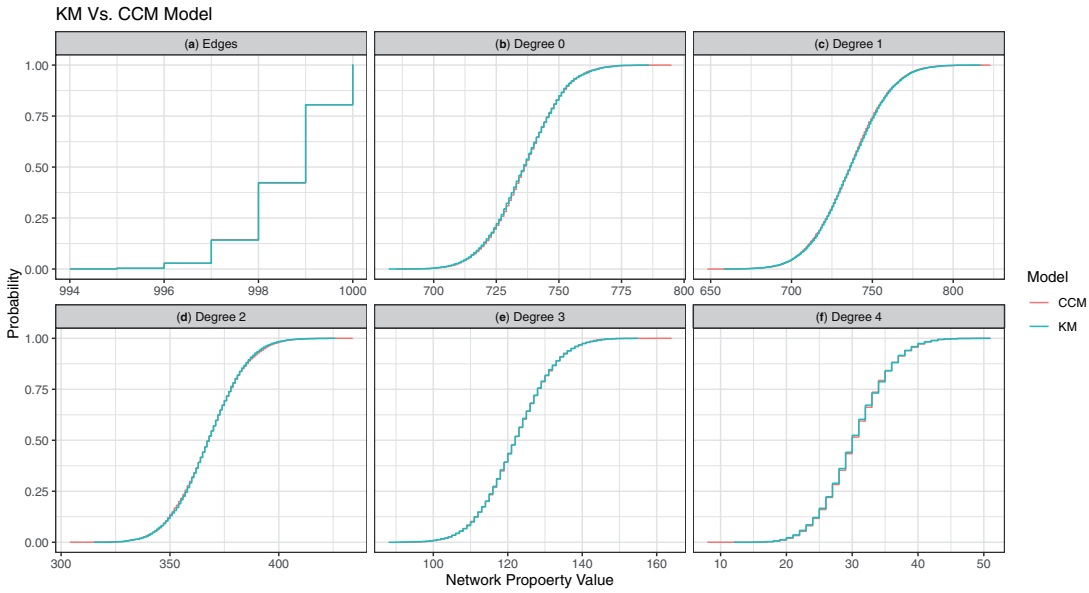

FIG. 4. Comparison between KM model and CCM on the number of edges and degree distribution: A comparison of the number of edges and number of nodes of specified degree across the network collection for the KM model and CCM ERGMs. Panel (a) depicts the CDF for the number of edges. Panels (b)–(f) depict the CDF for the number of nodes with degrees $\{0, 1, \ldots, 4\}$. The red lines depict the CDFs for the KM model and the blue lines depict the CDFs for the CCM. Because the CDFs match perfectly, the specified CCM appears to be able to capture the network structure generated by the KM model.

1. *Simulate the mechanistic model:* Let γ_s denote the microscopic rules associated with the serial monogamy setting for the KM model. We simulate $k = 10,000$ networks, $\{g_{M_1}, \ldots, g_{M_k}\}$, based on γ_s.
2. *Propose essential network property candidates:* Our candidate collection of essential network properties include only the number of individuals with degree 0. Let $X_{D_0}^{\gamma_r}$ represent the random variable for the number degree 0 nodes generated with γ_s.
3. *Estimate the joint probability distribution of essential network properties:* Let $P_{D_0}^{\gamma_s}$ denote the PMF for $X_{D_0}^{\gamma_s}$. We estimate $P_{D_0}^{\gamma_s}$, denoted as $\hat{P}_{D_0}^{\gamma_s}$, by letting $\hat{P}_{D_0}^{\gamma_s}(X_{D_0}^{\gamma_s} = x)$ equal the fraction of the k generated networks that have x individuals with degree 0.
4. *Sample networks:* We sample 10,000 networks based on the following PMF:

$$P_{\mathcal{G}}(G = g|\hat{P}_{D_0}^{\gamma_s}(X_{D_0}^{\gamma_s} = x)) \propto \left(\frac{1}{|c_{\eta(g)}|}\right) P_{\mathcal{C}}^{\gamma_s}(c_{\eta(g)}), \tag{5.3}$$

   where $P_{\mathcal{C}}^{\gamma_s}(c_{\eta(g)}) = \hat{P}_{D_0}^{\gamma_s}(X_{D_0}^{\gamma_s} = \eta(g))$ and $\eta(g)$ is the number of nodes with degree 0.
5. *Compare networks:* We compare the networks generated from the KM model to those generated from the CCM based on Equation (5.3) on a large set of network properties, which consists of number of edges, number of nodes of degrees 0 and 1 (nodes of higher degree are not compatible with the monogamy model) and eigenvalue centrality (max, mean, median and min across all nodes).

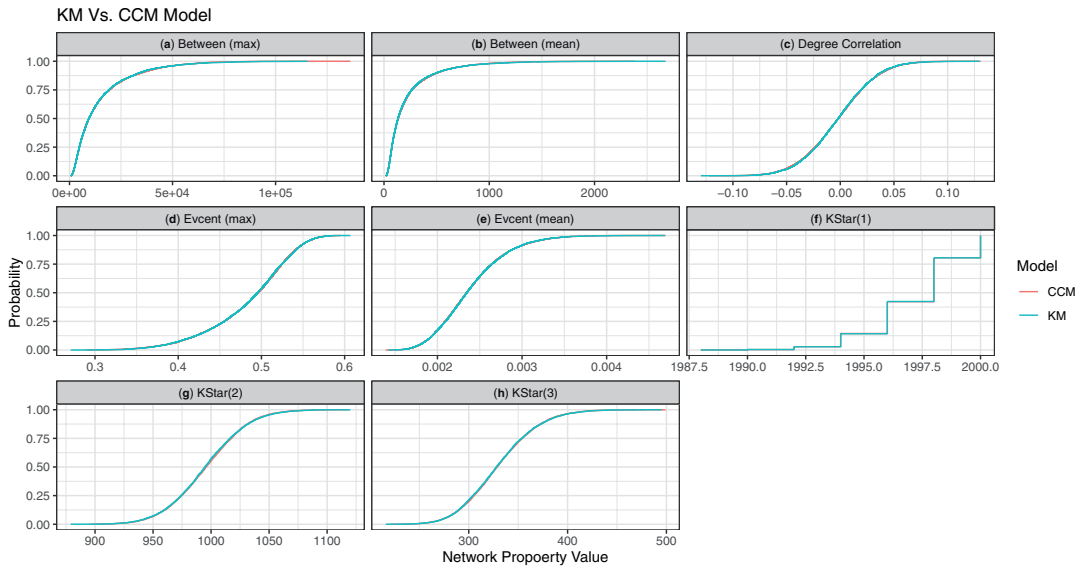

FIG. 5. Comparison between KM model and CCM on higher-order properties: A comparison of centrality measures (betweenness and eigenvector), degree correlation and number of k-stars across the network collection for KM model and CCM. Panels (a) and (b) depict the CDF for the max and mean betweenness centrality. Panel (c) depicts the CDF for the degree correlation. Panels (d) and (e) depict the CDF for the max and mean eigenvector centrality. Panels (f)–(h) depict the CDF for the number of k-stars with k equal to 1, 2 and 3. The red lines depict the CDFs for the KM model and the blue lines depict the CDFs for the CCM. Because the CDFs match perfectly, the specified CCM appears to be able to capture the network structure generated by the KM model.

Based on the Kolmogorov–Smirnov test, one cannot reject the hypothesis that the network property distributions are identical (the p-values ranged from 0.96 to 1 across all of the network properties).

6. *Iterate:* Based on the Kolmogorov–Smirnov tests, we conclude that the number of individuals with degree 0 is the only essential network property and the probability distribution $P^{ys}_{D_0}$ characterizes the serial monogamy KM mechanistic model.

Note that as individuals either have degree 0 or 1, it would be equivalent to use the number of individuals of degree 1 as our essential network property.

6. Limitations of CCMs

To our knowledge, all PNMs formulate the probability for a network g, $P_{\mathscr{G}}(G = g)$, based on the frequencies with which that particular network structures are present in g. For example, CCMs assume that all networks within a congruence class have the same probability [40]; this assumption is also made in commonly used PNMs including the ER model [20], stochastic block (SB) model [41] and ERGMs [22]. While ER and SB models are developed to model specific network structures—the number of edges and this number stratified by categorical–nodal covariates, respectively—CCMs and ERGMs do not have this constraint. In theory, CCMs and ERGMs can represent any PMF on the space of networks by including a sufficient number of parameters associated with network structures; but the number of parameters may be large and describe structures that consists of many nodes. Therefore, all MNMs—to the authors'

knowledge—can be represented as a CCM or an ERGM. However, there are practical constraints on these models. In the case of ERGMs, when the model includes network structures that are larger than a small number of nodes (e.g., 2–4), the parameter estimates will have difficulty converging [42, 43]. Furthermore, as demonstrated in Section 3, there are limitations with representing MNMs parsimoniously using ERGMs; it is not clear to the authors what ERGM terms are necessary to include to represent the KM model. For CCMs, the ratio of the size of two adjacent congruence classes must be evaluated to use an MCMC to generate networks from the model [33], which we expect to be increasing difficult as the size of the network structure increases. Below we provide an example of a mechanistic model that requires a PNM to incorporate a network structure that consists of a large number of nodes, which is difficult for both CCMs and ERGMs.

MNMs represent a range of phenomena in social and biological systems. Such phenomena include the small-world property, which refers to the notion that pairwise shortest path lengths are small—logarithmic in n (the number of nodes)—in most networks. The small-world property allows infections to potentially reach any individual in a population over relatively short transmission chains. Another common phenomenon in social systems is the coalescence of influence to a few individuals [44, 45]. This macroscopic phenomenon has been shown to emerge from a small collection of microscopic rules that encourage preferential attachment—the process wherein a new node introduced to the system links adjoins to an existing node with a probability proportional to the number of edges the node already has, i.e., its degree. Preferential attachment mechanistic rules were introduced in the model of Price for directed networks to study patterns of citation of scientific papers [44], and were later introduced independently in a different formulation for undirected networks by Barabási–Albert (BA) to describe a broad range of scientific and societal systems [45].

The BA model can be initiated with a small seed network, which grows by the addition of new nodes one at a time. (The model can be modified in various ways, but we consider only the original version of the model.) Nodes and edges, once introduced, are never deleted. Each new node forms exactly m_0 new edges with existing nodes based on the linear preferential attachment rule. One feature of the BA model is that it creates networks with a single connected component. Neither ERGMs nor CCMs can be formulated to handle this constraint, as both frameworks are only able to include model parameters associated with network structures that consist of only a small number of nodes (e.g., dyads, triangles and four-cycles).

The properties of the BA model illustrate the inability of current PNMs to represent certain popular mechanistic models. This is particularly striking given that the preferential attachment mechanistic rule for the BA model is designed to be straightforward. While the formulation of the preferential attachment rule clearly indicates its impact on the degree distribution of the resulting network, the impact of the rule on other networks features is less immediately apparent. For example, as discussed above, the rule leads to global features of the generated networks (e.g., presence of a single connected component). The preferential attachment rule has been observed to have an impact on other network features, such as correlations between the degrees of connected nodes [46] and network clustering coefficient [47].

To assess which mechanistic models are compatible with current PNMs requires an approach to quantifying the size of the network structures impacted by a mechanistic rule; for example, the approach would need to assess whether a rule impacts probabilities associated with pairs (size 2 structure), triads (size 3) or larger network structures. For the BA model, the rule impacts the global features of generated networks (i.e., size is equal to n). The formulation of an approach to quantifying size would enable creating a taxonomy or classification for mechanistic rules based on the size of the network structures impacted by the rule. It may be possible to assess whether the mechanism impacts network structures of small sizes (two–four nodes) by converting the mechanistic model to a probabilistic model using the

approach proposed in this article. However, even for these settings, there could be practical limitations on the development of a taxonomy via this process, as no straightforward method is available to assess the similarity of two mechanistic rules; hence, each rule would need to be assessed independently of other rules.

7. Discussion

In this article, we proposed the MPMC framework for first learning the joint distribution of essential network properties of a MNM and then using a probabilistic model, the CCM, to generate collections of networks that are indistinguishable from those generated by the original mechanistic model. There are advantages to being able to specify $P_{\mathscr{G}}(G = g|\gamma)$ as doing so enables investigators to perform statistical inference. In particular, the framework enables the investigation of whether a certain network property, such as clustering, is over- or under-represented in the generated networks compared with a reference model. There has been extensive research linking the presence or frequency of network properties to the nature of processes operating on networks, such as disease propagation [14–19]. MPMC also enables statistical testing of hypothesis, such as whether two distinct rules, γ^1 and γ^2, generate identical networks, i.e., whether $P_{\mathscr{G}}(G = g|\gamma^1) = P_{\mathscr{G}}(G = g|\gamma^2)$ for all $g \in \mathscr{G}$ or whether systems generated under two distinct rules have the same set of essential network properties.

An illustration of two examples of mechanistic models that are based on relatively simple rules demonstrates the complexity that can arise even from simple mechanistic models. This complexity helps to reveal the limitations on representing mechanistic models using probabilistic models. We identified three promising areas of future research. The first investigates ways to propose mechanistic models that are consistent with probability distributions of networks—the reverse of what we discussed above; see [48] for initial work in this area. The second is finding ways to allow the flexibility that probabilistic models require to represent a broader class, or particular classes, of mechanistic models. The third is a need for a taxonomy or classification of mechanistic models that is based on the set of network properties that are influenced by the mechanistic rules.

Our examples were kept simple in order to demonstrate a proof of concept of the framework; we acknowledge that much additional work is needed to bridge the two approaches. Nevertheless, the proposed framework provides a novel method for revealing relationships between the two approaches for modelling networks; this development has the potential to provide investigators with new insights about how networks are formed and how they impact the processes that operate on them.

Funding

This research is supported by the following grants from the National Institutes of Health: R37AI51164, R01AI138901, AI036214 and R01AI147441.

References

1. Boily, M.-C., Mâsse, B., Alsallaq, R., Padian, N. S., Eaton, J. W., Vesga, J. F. & Hallett, T. B. (2012) HIV treatment as prevention: Considerations in the design, conduct, and analysis of cluster randomized controlled trials of combination HIV prevention. *PLoS Med.*, **9**, e1001250.
2. Djurisic, S., Rath, A., Gaber, S., Garattini, S., Bertele, V., Ngwabyt, S.-N., Hivert, V., Neugebauer, E. A., Laville, M., Hiesmayr, M., Demotes-Mainard, J., Kubiak, C., Jakobsen, J. C. & Gluud, C. (2017) Barriers to the conduct of randomised clinical trials within all disease areas. *Trials*, **18**, 1–10.
3. Nichol, A., Bailey, M., Cooper, D. (2010) Challenging issues in randomised controlled trials. *Injury*, **41**, S20–S23.

4. Pearce, W., Raman, S. & Turner, A. (2015) Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy. *Trials*, **16**, 1–7.

5. Cai, X., Loh, W. W. & Crawford, F. W. (2021) Identification of causal intervention effects under contagion. *J. Causal Inference*, **9**, 9–38.

6. Ogburn, E. L. & VanderWeele, T. J. (2014) Causal diagrams for interference. *Stat. Sci.*, **29**, 559–578.

7. Carnegie, N. B., Wang, R. & De Gruttola, V. (2016) Estimation of the overall treatment effect in the presence of interference in cluster-randomized trials of infectious disease prevention. *Epidemiol. Methods*, **5**, 57–68.

8. Wang, R., Goyal, R., Lei, Q., Essex, M. & De Gruttola, V. (2014) Sample size considerations in the design of cluster randomized trials of combination HIV prevention. *Clin. Trials*, **11**, 1740774514523351.

9. Halloran, M. E. & Hudgens, M. G. (2016) Dependent happenings: a recent methodological review. *Curr. Epidemiol. Rep.*, **3**, 297–305.

10. Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B. George, L., Jastrzębski, M., Izzo, A. S., Fowler, G., Palmer, A., Delport, D., Scott, N., Kelly, S. L., Bennette, C. S., Wagner, B. G., Chang, S. T., Oron, A. P., Wenger, E. A., Panovska-Griffiths, J., Famulare, M. & Klein, D. J. (2021) Covasim: an agent-based model of COVID-19 dynamics and interventions. *PLoS Comput. Biol.*, **17**, e1009149.

11. Hinch, R., Probert, W. J., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Bulas Cruz, A., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Legat, O., Bentley, K., Mead, T., Van-Vuuren, K., Feldner-Busztin, D., Ristori, T., Finkelstein, A., Bonsall, D. G., Abeler-Dörner, L. & Fraser, C. (2021) Open ABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS Comput. Biol.*, **17**, e1009146.

12. Goyal, R., Hotchkiss, J., Schooley, R. T., De Gruttola, V., Martin, N. K. et al. (2021) Evaluation of SARS-CoV-2 transmission mitigation strategies on a university campus using an agent-based network model. *Clin. Infect. Dis.*, **73**, 1735–1741.

13. Hambridge, H. L., Kahn, R. & Onnela, J.-P. (2021) Examining SARS-CoV-2 interventions in residential colleges using an empirical network. *Int. J. Infect. Dis.*, **113**, 325–330.

14. Newman, M. (2002) Assortative mixing in networks. *Phys. Rev. Lett.*, **89**, 208701.

15. Newman, M. E. (2003a) Mixing patterns in networks. *Phys. Rev. E*, **67**, 026126.

16. Boguná, M., Pastor-Satorras, R. & Vespignani, A. (2003) Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.*, **90**, 028701.

17. Newman, M. E. (2003b) Properties of highly clustered networks. *Phys. Rev. E*, **68**, 026121.

18. Keeling, M. J. & Eames, K. T. D. (2005) Networks and epidemic models. *J. R. Soc. Interface*, **2**, 295–307.

19. Miller, J. C. (2009) Percolation and epidemics in random clustered networks. *Phys. Rev. E*, **80**, 020901.

20. Erdős, P. & Rényi, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 17–61.

21. Frank, O. & Strauss, D. (1986) Markov graphs. *J. Am. Stat. Assoc.*, **81**, 832–842.

22. Robins, G., Pattison, P., Kalish, Y. & Lusher, D. (2007) An introduction to exponential random graph (p*) models for social networks. *Soc. Netw.*, **29**, 173–191.

23. Kretzschmar, M. & Morris, M. (1996) Measures of concurrency in networks and the spread of infectious disease. *Math. Biosci.*, **133**, 165–195.

24. Morris, M. & Kretzschmar, M. (1997) Concurrent partnerships and the spread of HIV. *AIDS*, **11**, 641–648.

25. Morris, M., Goodreau, S., & Moody, J. (2008) Sexual networks, concurrency, and STD/HIV. *Sex. Transm. Dis.*, **4**, 109–125.

26. Palombi, L., Bernava, G. M., Nucita, A., Giglio, P., Liotta, G., Nielsen-Saines, K., Orlando, S., Mancinelli, S., Buonomo, E., Scarcella, P., Altan, A. M. D., Guidotti, G., Ceffa, S., Haswell, J., Zimba, I., Magid, N. A. & Marazzi, M. C. (2012) Predicting trends in HIV-1 sexual transmission in sub-Saharan Africa through the Drug Resource Enhancement Against AIDS and Malnutrition model: antiretrovirals

for reduction of population infectivity, incidence and prevalence at the district level. *Clin. Infect. Dis*., **55**, 268–275.

27. Lusher, D., Koskinen, J. & Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications.* Cambridge University Press.

28. Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B (Methodol.)*, **36**, 192–225.

29. Pressé, S., Ghosh, K., Lee, J. & Dill, K. A. (2013) Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, **85**, 1115.

30. Newman, M. E. (2010) *Networks an Introduction*. New York: Oxford University Press.

31. Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A. & Caldarelli, G. (2019) The statistical physics of real-world networks. *Nat. Rev. Phys.*, **1**, 58.

32. Molloy, M. & Reed, B. (1995) A critical point for random graphs with a given degree sequence. *Random Struct. Algor.*, **6**, 161–180.

33. Goyal, R., Blitzstein, J. & De Gruttola, V. (2014) Sampling networks from their posterior predictive distribution. *Netw. Sci.*, **2**, 107–131.

34. Goyal, R. & De Gruttola, V. (2018) Inference on network statistics by restricting to the network space: applications to sexual history data. *Stat. Med.*, **37**, 218–235.

35. Goyal, R. & De Gruttola, V. (2020) Dynamic network prediction. *Netw. Sci.*, **8**, 574–595.

36. Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, vol. 26. Cambridge, UK: CRC Press.

37. Scott, D. W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

38. Wasserman, S. & Faust, K. (1994) *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge University Press.

39. Arnold, T. B. & Emerson, J. W. (2011) Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, **3**, 34–39.

40. Goyal, R., Carnegie, N., Slipher, S., Turk, P., Little, S. J. & De Gruttola, V. (2023) Estimating contact network properties by integrating multiple data sources associated with infectious diseases. *Stat. Med*., **42**, 3593–3615.

41. Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Soc. Netw.*, **5**, 109–137.

42. Blackburn, B. & Handcock, M. S. (2022) Practical network modeling via tapered exponential-family random graph models. *J. Comput. Graph. Stat.*, **32**, 1–14.

43. Schweinberger, M. (2011) Instability, sensitivity, and degeneracy of discrete exponential families. *J. Am. Stat. Assoc.*, **106**, 1361–1370.

44. Price, D. d. S. (1976) A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inform. Sci.*, **27**, 292–306.

45. Barabasi, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

46. Qu, J., Wang, S.-J., Jusup, M. & Wang, Z. (2015) Effects of random rewiring on the degree correlation of scale-free networks. *Sci. Rep.*, **5**, 15450.

47. Klemm, K. & Eguiluz, V. M. (2002) Growing scale-free networks with small-world behavior. *Phys. Rev. E*, **65**, 057102.

48. Chen, S., Mira, A. & Onnela, J.-P. (2020) Flexible model selection for mechanistic network models. *J. Complex Netw.*, **8**, cnz024.