METRON (2023) 81:21–35
https://doi.org/10.1007/s40300-023-00246-3

# Modeling of networked populations when data is sampled or missing

Ian E. Fellows[1] · Mark S. Handcock[2]

Received: 3 October 2022 / Accepted: 22 March 2023 / Published online: 20 May 2023

## Abstract

Networked populations consist of inhomogeneous individuals connected via relational ties. The individuals typically vary in multivariate attributes. In some cases primary interest focuses on individual attributes and in others the understanding of the social structure of the ties. In many circumstances both are of interest, as is their relationship. In this paper we consider this last, most general, case. We model the joint distribution of social ties and individual attributes when the population is only partially observed. Of central interest is when the population is surveyed using a network sampling design. A second situation is when data about a subset of the ties and/or the individual attributes is unintentionally missing. Exponential-family random network models (ERNM)s are capable of specifying a joint statistical representation of both the ties of a network and individual attributes. This class of models allow the nodal attributes to be modeled as stochastic processes, expanding the range and realism of exponential-family approaches to network modeling. In this paper we develop a theory of inference for ERNMs when only part of the network is observed, as well as specific methodology for partially observed networks, including non-ignorable mechanisms for network-based sampling designs. In particular, we consider data collected via contact tracing, of considerable importance to infectious disease epidemiology and public health.

**Keywords** Survey methods · Social networks · Missing data · Network sampling · Exponential families · Contact tracing · Epidemic modeling

## 1 Introduction

It is not uncommon for researchers to collect data on a subset of a single network rather than observing the full network. This partially observed case has been studied within the framework of exponential-family random graph models (ERGM) by [11, 13]. However their

✉ Ian E. Fellows
ian@fellstat.com

Mark S. Handcock
handcock@stat.ucla.edu

[1] Fellows Statistics, San Diego, CA 92107, USA

[2] Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095-1554, USA

formulation suffers from the limitation that any nodal attributes included in the model must be fully observed, and only ties may be partially observed. This assumption is not met in most sampling designs, where only some of the nodes are surveyed by the researcher, and reduces the practical usage of ERGMs in the survey sampling and missing data setting.

By including nodal attributes as stochastic variates rather than fixed quantities, exponential-family random network models [3, ERNM] can provide a convenient basis for inference in cases where the data is partially unobserved, either due to designed, or out-of-design (e.g., survey non-response) mechanisms.

While our framework is applicable to all partial observation mechanisms we consider two common mechanisms for partial observations in more detail, specifically:

Missing data: If the population is comprised of a large number of units/nodes, or the number of ties is large, it is relatively common to find that the resources to observe a full network are not available. Often nodal or tie variables are unavailable for sampling or do not provide complete responses to a survey instrument. In this case, only some of the tie variables and nodal attributes are collected. We treat missing data as a form of sampling in which the sampling mechanism is unknown and outside the control of the researcher, or an *out-of-design missing data mechanism*. A good example of this is the National Longitudinal Study of Adolescent Health (Add Health), a school-based, longitudinal study of the health-related behaviors of adolescents and their outcomes in young adulthood. The study design sampled 80 high schools and 52 middle schools from the U.S., representative with respect to region of country, urbanicity, school size, school type, and ethnicity [16]. In 1994–95 an in-school questionnaire was administered to a nationally representative sample of students in grades 7 through 12. In addition to demographic and contextual information, each respondent was asked to nominate up to five boys and five girls within the school whom they regarded as their best friends. Thus each student could nominate up to ten students within the school [26]. The nominations and contextual information were not available for some of the adolescents, either due to absence from school while the survey was being conducted, or refusal to participate. Thus, both the tie and nodal variates contained missing values.

Network sampling designs: Many studies in hard-to-reach populations use study designs that trace the linkages of an underlying social network. The designs lead to partial observation of the attributes and ties of the networked population. Note that the tie variables are often of secondary interest, although they are what makes it possible to conduct the survey. Such sampling designs have been exploited to estimate population disease rates and population sizes [8–10, 14, 15].

In this paper we develop approaches for each of these scenarios in the context of ERNMs. Sections 2 through 4 introduce ERNM and extend the theory to incorporate partially observed populations. Section 5 develops methodology for each of the scenarios. Section 5.1 looks at the effect of random non-response on a real data-set. Section 5.2 considers sampling designs where the seeds are not chosen at random from the population, but are some form of convenience sample chosen by an at least partially unknown process. Section 5.3 develops estimates based on contact tracing designs, which is of vital importance to the public health community. To our knowledge, the methods outlined in this paper represent the first statistically justifiable approach to inference from contract tracing data.

## 2 Exponential-family random network models

Exponential-family random network models [3] are a generalization of the exponential-family random graph model [7, 18], where both tie variables and nodal attributes are treated as random variates. Formally, in a population of $n$ units, let $Y_{i,j}$ indicate that unit $i$ has a tie to unit $j$. Let $Y$ be an $n \times n$ matrix $[Y_{i,j}]$ and $X$ be a an $n \times K$ matrix $[X_{ik}]$ of unit attributes. We define a *network* $Z$ as the union of the nodal attributes and the *graph* structure (i.e. $Z = \{X, Y\}$). An exponential-family model of $Z$ is expressed as

$$p(Z = z \mid \eta) = \frac{1}{c(\eta, \mathcal{Z})} e^{\eta \cdot g(z)} \quad t \in \mathcal{Z}, \tag{1}$$

where $\eta \in R^q$ is a vector of parameters, $g$ is a $q-$vector valued function defining a set of sufficient statistics, $\mathcal{Z}$ is the sample space of networks and $c(\eta, \mathcal{Z}) = \int_{z \in \mathcal{Z}} e^{\eta \cdot g(z)}$ is the normalizing constant. This model is developed in [3]. We focus on the case where the relational ties are binary and the sample space of each covariate is countable. The case of valued ERGM and ERNM is a straight-forward extension, and we focus on the binary/countable case to explicate the partial observational aspects and their impacts on inference.

### 2.1 The simple homophily model

Though any set of network statistics can be represented by $g$ in Eq. (1) [3], the analysis in this paper will focus on a particularly parsimonious, but powerful, network model. Suppose that $X = (X_1, \ldots, X_n)$ is a univariate categorical variable with $m$ levels, labeled $0, \ldots, m-1$. If $X_i = l$ we say that unit $i$ is in group $l$. A joint model for $X$ and $Y$ is

$$p(Z = (y, x) \mid \eta) = \frac{1}{c(\eta, \mathcal{Z})} e^{\eta_0 \sum_{i,j} y_{i,j} + \eta_2 h(y,x) + \sum_{l=0}^{m-2} \eta_{j+3} \sum_{i=1}^{n} I(x_i = l)} \quad (y, x) \in \mathcal{Z}.$$

The first term of this model is the number of ties, and controls the density of the graph. The last term represents the number of nodes in each category of $x$, except for the last level, which is dropped to maintain identifiability of the model. The second term $h$ is the regularized sample homophily of $x$, as introduced by [3], and is defined as

$$h(y, x) = \sum_{k=0}^{m-1} \sum_{i : x_i = k} \sqrt{d_{i,k}(y, x)} - E_{\perp\!\!\!\perp}\left(\sqrt{d_{i,k}(y, x)}\right),$$

where $d_{i,k}(y, x)$ is the number of ties between node $i$ and nodes in group $k$, and $E_{\perp\!\!\!\perp}(f(Y, X))$ is the expectation of the statistic $f(Y, X)$, conditional upon $Y = y$ and the category counts (that is, the number of nodes in each category of $x$, $n(x) = \{n_k(x)\}_{k=1}^{K}$), assuming that $X$ and $Y$ are independent. Thus, each term in the sum is the square root of the number of neighbors of a node which share the same category, minus what would be expected by chance. Using this form of homophily avoids the degeneracy problems found in other formulations. For a more thorough justification, see [3].

While the examples in this paper focus on applications of the simple homophily model, the framework presented here applies to any arbitrary set of network statistics $g$. For example, in many applications the nodal attributes are multivariate, and their relationships are of interest to the researcher. Fellows and Handcock [3] developed a network statistic that can be interpreted as a conditional logistic regression term which, if included, can model the relationship of several categorical variates.

## 3 Likelihood-based inference from partially observed networks

In this section we develop likelihood-based inference for network models based on partial observation of the networks. The approach allows non-ignorable sampling mechanisms for the networks, including some common network-based sampling designs.

References [11, 13] developed a theory of partially observed networks for ERGMs, and the specification for ERNMs proceeds similarly, though our formulation supports a more general class of partial observation processes known as *amenable* [13] or *missing not at random* [22, MNAR; see]. Let $Z_{obs}$ and $Z_{miss}$ represent, respectively, the observed and unobserved part of the complete network $Z$. We write $Z = (Z_{obs}, Z_{miss})$, with realizations $z = (z_{obs}, z_{miss})$. Let $W$ be a random variable representing the sampling process with realization $w$. The probabilistic distribution of $W$ is the *sampling mechanism*, and must fully specify the sample selection process, including the partition of $Z$ into $Z_{obs}$ and $Z_{miss}$. Typically, $W$ will consist of an $n$ by $n$ matrix, $[W_{ij}]$, indicating whether the $(i, j)$th dyad was sampled, and an $n$ by $K$ matrix indicating which nodal attributes are missing, not sampled or otherwise not observed. However, $W$ may contain additional information about the partial observation process, such as the order of observation.

We write the *full data likelihood* as

$$p(Z = z, W = w \mid \eta, \theta) = p(W = w \mid Z = z, \theta)\frac{1}{c(\eta, \mathcal{Z})}e^{\eta \cdot g(z)},$$

and we wish to draw inferences about $\eta$ from the *observed data likelihood*, defined as

$$\begin{aligned} p(Z_{obs} &= z_{obs}, W = w \mid \eta, \theta) \\ &= \sum_{z_{miss}} p(W = w \mid z = (z_{obs}, z_{miss}), \theta)\frac{1}{c(\eta, \mathcal{Z})}e^{\eta \cdot g((z_{obs}, z_{miss}))}. \end{aligned} \tag{2}$$

This probability model jointly represents the distribution of the network $Z$, and the sampling process $W$. The functional form of $p(W = w \mid Z = z, \theta)$ is dependent on the form of partial observation, and will differ depending on how $Z_{obs}$ was obtained. Section 5.2 illustrates a design of particular interest known as *biased seed link tracing*. When the sampling probabilities only depend on the observed data, the sampling design is *amenable* to the model [13], and is ignorable in the sense of [22]. In this case, the observed data likelihood simplifies to

$$\begin{aligned} p(Z_{obs} &= z_{obs}, W = w \mid \eta, \theta) \\ &= \sum_{z_{miss}} p(W = w \mid Z_{obs} = z_{obs}, \theta)\frac{1}{c(\eta, \mathcal{Z})}e^{\eta \cdot g((z_{obs}, z_{miss}))} \\ &= p(W = w \mid Z_{obs} = z_{obs}, \theta)\sum_{z_{miss}} \frac{1}{c(\eta, \mathcal{Z})}e^{\eta \cdot g((z_{obs}, z_{miss}))} \\ &\propto \sum_{z_{miss}} \frac{1}{c(\eta, \mathcal{Z})}e^{\eta \cdot g((z_{obs}, z_{miss}))}. \end{aligned} \tag{3}$$

Thus, when the sampling process is ignorable, inferences on $\eta$ are not affected by $p(W = w \mid Z_{obs} = z_{obs}, \theta)$, and so knowledge of the sampling process is not essential for the process of likelihood-based inference.

Having defined the full and observed data likelihood, it is also useful to define the complementary *missing data likelihood*:

$$\begin{aligned} p(Z_{miss} = z_{miss} \mid W = w, Z_{obs} = z_{obs}, \eta, \theta) \\ = \frac{p(W = w \mid Z = (z_{obs}, z_{miss}), \theta) e^{\eta \cdot g((z_{obs}, z_{miss}))}}{c(z_{obs}, w, \eta, \theta)} \end{aligned} \tag{4}$$

where

$$c(z_{obs}, w, \eta, \theta) = \sum_{z_{miss}} p(W = w \mid Z = (z_{obs}, z_{miss}), \theta) e^{\eta \cdot g((z_{obs}, z_{miss}))}.$$

The (observed data) likelihood can then be rewritten as the ratio of two normalizing constants

$$\begin{aligned} & p(Z_{obs} = z_{obs}, W = w \mid \eta, \theta) \\ & \quad = \frac{1}{c(\eta, \mathcal{Z})} \sum_{z_{miss}} p(W = w \mid Z = (z_{obs}, z_{miss}), \theta) e^{\eta \cdot g((z_{obs}, z_{miss}))} \\ & \quad = \frac{c(z_{obs}, w, \eta, \theta)}{c(\eta, \mathcal{Z})}, \end{aligned}$$

and using this, we may write the observed data log-likelihood ratio of $(\eta, \theta)$ versus $(\eta_0, \theta_0)$ as

$$\begin{aligned} \ell(\eta, \theta) - \ell(\eta_0, \theta_0) &= \log\left( \frac{c(z_{obs}, w, \eta, \theta)}{c(z_{obs}, w, \eta_0, \theta_0)} \right) - \log\left( \frac{c(\eta, \mathcal{Z})}{c(\eta_0, \mathcal{Z})} \right) \\ &= \log\left( \sum_{z_{miss}} \frac{p(W = w \mid Z = z, \theta)}{p(W = w \mid Z = z, \theta_0)} e^{(\eta - \eta_0) \cdot g(z)} \frac{p(W = w \mid Z = z, \theta_0) e^{\eta_0 \cdot g(z)}}{c(z_{obs}, w, \eta_0, \theta_0)} \right) \\ &\quad - \log\left( \sum_{z_{miss}} e^{(\eta - \eta_0) \cdot g(z)} \frac{e^{\eta_0 \cdot g(z)}}{c(\eta, \mathcal{Z})} \right) \\ &= \log\left( E_{\eta_0, \theta_0} \left( \frac{p(W = w \mid Z, \theta)}{p(W = w \mid Z, \theta_0)} e^{(\eta - \eta_0) \cdot g(Z)} \right) \mid W = w, Z_{obs} = z_{obs} \right) \\ &\quad - \log\left( E_{\eta_0} (e^{(\eta - \eta_0) \cdot g(Z)}) \right) \\ &= \log\left( E_{\eta_0, \theta_0} (e^{(\eta - \eta_0) \cdot g(Z)} \mid Z_{obs} = z_{obs}) \right) - \log\left( E_{\eta_0} (e^{(\eta - \eta_0) \cdot g(Z)}) \right) \\ &\quad + \log\left( \frac{E_{\eta, \theta} \ (p(W = w \mid Z, \theta\ ) \mid Z_{obs} = z_{obs})}{E_{\eta_0, \theta_0} (p(W = w \mid Z, \theta_0) \mid Z_{obs} = z_{obs})} \right). \end{aligned} \tag{5}$$

Both Eqs. (4) and (5) motivate algorithms to draw inferences about $\eta$ and $\theta$. Section 4 describes the algorithm motivated by Eq. (4), and Appendix A.1 outlines an algorithm using Eq. (5).

## 4 Calculating the MLE with MCMC

For most models, Eq. (4) is not analytically solvable as the normalizing constant is computationally intractable. However we may approximate it by Markov Chain Monte Carlo (MCMC). Let $z^{(k)}$ and $z_m^{(k)}$ where $k \in (1, \ldots, M)$ be samples from the full data likelihood and missing data likelihood, respectively, with parameters $\eta_0, \theta_0$. Then Eq. (4) may be approximated by

$$\ell(\eta,\theta)-\ell(\eta_0,\theta_0)$$
$$\approx \log\left(\frac{1}{M}\sum_{k}^{M}\frac{p(w\mid z_m^{(k)},\theta)}{p(w\mid z_m^{(k)},\theta_0)}e^{(\eta-\eta_0)\cdot g(z_m^{(k)})}\right)-\log\left(\frac{1}{M}\sum_{k}^{M}e^{(\eta-\eta_0)\cdot g(z^{(k)})}\right) \quad (6)$$

For fixed $M$, as $\eta,\theta$ move away from $\eta_0,\theta_0$ the variance of this approximation increases (Note that it is essentially a Hájek estimator [24]). Because we will be optimizing Eq. (4), it is useful to have both the first and second derivatives of the log-likelihood, which are

$$\left[\frac{\delta\ell}{\delta\eta}\right]_i = E_{\eta,\theta}(g_i(t)\mid Z_{obs}=z_{obs},W=w)-E_{\eta,\theta}(g_i(Z))$$
$$\left[\frac{\delta^2\ell}{\delta\eta_i\delta\eta_j}\right]_{ij} = -\mathrm{cov}(g_i(Z),g_j(Z))$$
$$+\mathrm{cov}(g_i(Z),g_j(Z)\mid Z_{obs}=z_{obs},W=w)\quad i,j\in\{1,\dots,q\}.$$

The expectations and covariances in these derivatives can be approximated using the conditional and unconditional MCMC samples and thus we can then use the following algorithm to approximate the maximum likelihood estimator (MLE).

1. Let $l=0$ and choose initial parameter values $\eta^{(0)}$, $\theta_0$.
2. Use MCMC to generate $M_{\text{miss}}$ samples, $z_{miss}^{(k)}$ from $p(Z_{miss}=z_{miss}\mid\eta^l,Z_{obs}=z_{obs},W=w)$.
3. Use MCMC to generate $M_{\text{full}}$ samples $z^{(k)}$ from $p(Z=z\mid\eta^l)$.
4. Using the samples from Steps 2 and 3 in Eq. (6), find $\eta^{l+1},\theta^{l+1}$ maximizing the likelihood ratio, subject to the convergence condition $\left\|\eta^{l+1}-\eta^l\right\|<\epsilon$ and $\left\|\theta^{l+1}-\theta^l\right\|<\epsilon$.
5. If the likelihood has not converged, set $l=l+1$ and go to Step 2.
6. Let the MLE estimate be $\hat{\eta}=\eta^{l+1}$ and $\hat{\theta}=\theta^{l+1}$

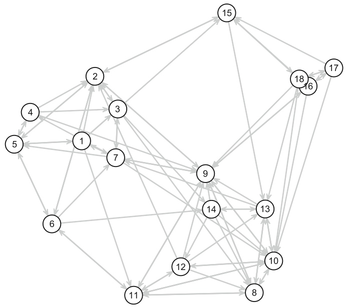
Asymptotic standard errors for $\hat{\eta}$ may be obtained using an MCMC approximation to the Fisher information matrix (i.e. the second derivative of the log-likelihood). While asymptotics of the Fisher information matrix are not assured with respect to ERNM (or ERGM) models, [3] show strong empirical agreement between the Fisher information-based standard errors and parametric bootstrap simulations. Standard errors for the mean value parameters $\hat{\mu}=E(g(Z)\mid\eta=\hat{\eta})$ can be approximated by MCMC sampling (as in Step 3).

## 5 Specific forms of partial observation

In this section we consider the three common forms of partial observation considered in the introduction, each corresponding to a different mechanism of partial observation or conceptualization of that mechanism.

### 5.1 Missing data: unobserved relational information

It is common when surveying networked populations that there are insufficient resources to conduct a census of the population and their relations. For efficiency reasons, a sampling based survey is undertaken, or the full network is partially observed due to non-response. In this sub-section, we give an illustration of the effect of non-response where the dyad



| | | | |
|---|---|---|---|
| 1 | Ramauld (L) | 10 | Gregory (T) |
| 2 | Bonaventure (L) | 11 | Hugh (T) |
| 3 | Ambrose (L) | 12 | Boniface (T) |
| 4 | Berthold (L) | 13 | Mark (T) |
| 5 | Peter (L) | 14 | Albert (T) |
| 6 | Louis (L) | 15 | Amand (O) |
| 7 | Victor (L) | 16 | Basil (O) |
| 8 | Winfred (T) | 17 | Elias (O) |
| 9 | John (T) | 18 | Simplicius (O) |

**Fig. 1** Relationships among monks within a monastery and their affiliations as identified by Sampson: Young (T)urks, (L)oyal Opposition, and (O)utcasts

information is missing completely at random. We consider the relations of “liking” among 18 monks in a monastery [23]. The network analyzed has a directed tie between two monks if the sender monk ranked the receiver monk in the top three monks for positive affection in any of the three interviews given over a twelve month period [17]. The sociogram of this data-set is shown in Fig. 1. One nodal attribute of interest is an indicator of attendance at the minor “Cloisterville" seminary before coming to the monastery.

We fit a simple homophily model on Cloisterville status using the full data. In this model the three statistics are the count of the number of ties, the homophily statistic $h(y, x)$ and the count of the number of monks who attended the minor “Cloisterville" seminary before coming to the monastery. The corresponding natural parameters are $\eta_1$, $\eta_2$ and $\eta_3$, respectively. The corresponding mean-value parameters are $\mu_1$, $\mu_2$ and $\mu_3$, respectively. We then ran simulations on the effect of missingness by selecting tie variables, and Cloisterville status variates, completely at random and setting them to missing. Figure 2 shows one simulated missingness pattern with 15% missing. We ran 100 simulations at each missingness percentage. We note that this amounts to 100 simulations times 60 conditions for a total of 6000 MLE fits being performed. Each MLE fit required a large number of MCMC chain updates to be performed in order to ensure converge. Means and standard deviations of the ERNMs fit to these simulated missingness patterns are displayed in Fig. 3.

We see that the standard deviations of the estimates increase as the amount of missingness increases. At the higher missingness levels a small amount of bias is apparent relative to the full data MLE, but not more than one standard deviation. Note that the estimates of the mean value parameters are conditionally unbiased given the observed data, but can be unconditionally biased. One possible explanation for this bias is that there were only six monks who attended Cloisterville, and so at 50% missingness, a significant number of samples will include no (or perhaps a single) Cloisterville monks. In this case there is no direct information on homogeneity.

### 5.2 Network sampling: biased seed link-tracing

Network sampling is the general idea of using the social ties within a networked population to help guide the selection of samples from that population [25]. Handcock and Gile [13] explored the idea of sampling networks by tracing the ties. As a general concept, *link tracing* involves selecting one or more *seed* nodes, and then observing the ties connected to those

| | Y | | | | | | | | | | | | | | | | | | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | |
| 1 | 🟥 | 🟨 | 🟨 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| 2 | 🟨 | 🟥 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| 3 | 🟨 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 NA | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 |
| 4 | 🟨 | 🟨 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟨 | 🟨 | 🟥 |
| 5 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟨 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟨 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟥 |
| 6 | 🟨 | 🟥 | 🟥 | 🟨 | 🟨 | 🟥 | 🟨 | 🟥 | 🟨 | 🟥 | 🟨 NA | 🟥 NA | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 |
| 7 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 | 🟨 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| 8 | 🟥 | 🟥 | 🟥 NA | 🟥 NA | 🟥 NA | 🟥 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 | 🟥 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 |
| 9 | 🟨 | 🟨 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| 10 | 🟨 | 🟥 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 NA | 🟨 | 🟨 | 🟥 | 🟨 | 🟨 | 🟥 | 🟥 | 🟨 | 🟥 | 🟨 | 🟥 | 🟥 |
| 11 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 NA | 🟥 | 🟥 | 🟥 NA |
| 12 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 NA | 🟥 NA | 🟥 | 🟨 NA | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 NA | 🟥 | 🟨 | 🟨 |
| 13 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟨 | 🟥 NA | 🟥 | 🟨 NA | 🟥 | 🟥 |
| 14 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟨 | 🟨 | 🟥 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 |
| 15 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 | 🟨 NA | 🟥 | 🟥 | 🟥 NA | 🟥 NA | 🟨 NA |
| 16 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 |
| 17 | 🟥 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟨 | 🟨 | 🟨 | 🟨 NA | 🟥 | 🟨 | 🟨 NA |
| 18 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 NA | 🟨 | 🟥 NA | 🟨 | 🟥 NA | 🟥 | 🟥 | 🟥 | 🟨 |

**Fig. 2** A variant of Sampson's monk's with 15% missingness. The Cloisterville status variable is marked on the right hand side. The cells with a present tie are indicated in red and those with absent ties are in pale yellow. The NA represent "not available" values (that is, missing values) (color figure online)

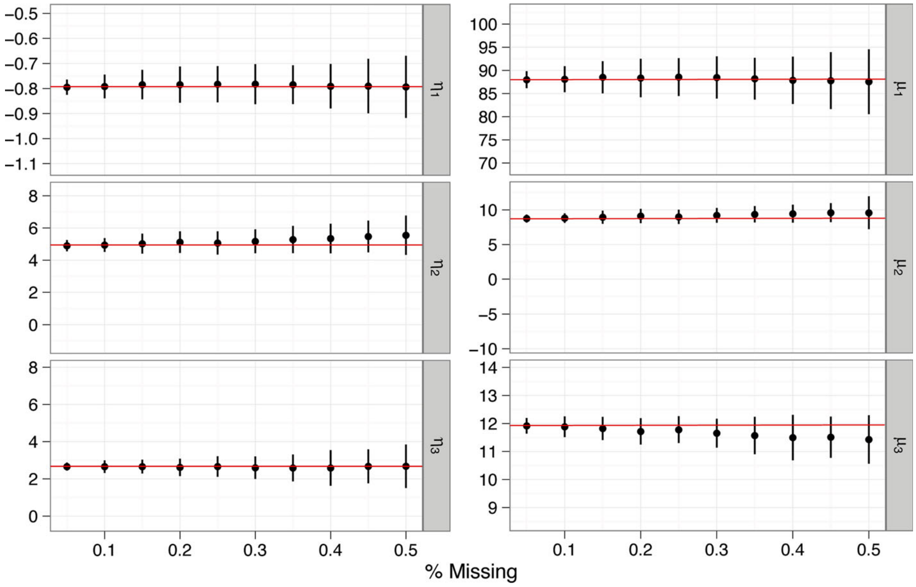


**Fig. 3** Means and standard deviations of model estimates of the natural and mean-value parameters. The means are represented by the black points. The magnitudes of the standard deviations are represented by the black lines. Red lines indicate fully observed MLE (color figure online)

seeds. One or more of these ties are then followed to the neighboring node, whose ties are observed, and the process is continued. Each iteration of this process is known as a *wave*.

Provided that the seed nodes are chosen at random, and the method by which ties are chosen to be followed depends only on the observed data, this missingness process is ignorable. To be explicit, consider a link tracing process with $k$ waves. Let $w_i$ be the ordered set of nodes and ties sampled in the $i$th wave in the order in which they were sampled, $w = \{w_0, \ldots, w_k\}$, and $w_{-i} = \{w_0, \ldots, w_{i-1}, w_{i+1}, \ldots, w_k\}$. If the seeds are chosen at random, and the ties followed by the sampling process are also chosen at random, then $p(W = w \mid Z = z, \theta) = p(W = w \mid Z_{obs} = z_{obs}, \theta)$, implying that the missingness is ignorable.

In many cases, however, the seeds are not chosen at random from the population, but are some form of convenience sample chosen by an at least partially unknown process. For example, in a population where some people have an infection and others do not, we may start with a sample of $s_{\text{inf}}$ seeds picked at random from among the infected individuals, and $s_{-\text{inf}}$ seeds picked from the non-infected individuals. These seeds are then used as a starting point for standard link tracing. We may then write the sampling probability as

$$\begin{aligned} p(w \mid t, \theta) &= p(w_0 \mid t, \theta) p(w_{-0} \mid z_{obs}, w_0, \theta) \\ &= \frac{(n_{\text{inf}} - s_{\text{inf}})!}{n_{\text{inf}}} \frac{(n_{-\text{inf}} - s_{-\text{inf}})!}{n_{-\text{inf}}!} p(w_{-0} \mid z_{obs}, w_0, \theta), \end{aligned}$$

where $n_{\text{inf}}$ and $n_{-\text{inf}}$ are the number of infected and non-infected in the population, respectively. Note that $p(w_{-0} \mid z_{obs}, w_0, \theta)$ does not depend on $z_{miss}$ and may be factored out of the likelihood in Eq. (2). Thus there is no need to calculate or model $p(w_{-0} \mid z_{obs}, w_0, \theta)$ explicitly, as it makes no impact on the likelihood. Hence, in this case, we can compute the likelihood without knowing the specific mechanism of seed selection.

### 5.3 Network sampling: positive contact tracing

As emerging epidemics develop, control measures (e.g., treatment, isolation and culling) focus on those members of the population that are known to have the infection. Because there are often many infected people who are unobserved, public health control can be ineffective (e.g., HIV [20] and COVID-19 [6]). The alternative of applying control measures to the entire population can be economically infeasible or ineffective (e.g., some instances of safe sex education, early stage COVID-19) [19, 20]. Contact tracing is the hybrid approach of treating both the known infected individuals and those who may have been infected by them [19, 20]. In U.S. public health, health clinics are required by state law to notify those at risk from infection due to their sexual relations with individuals tested, and found to be infected, by the clinic. The process of locating, notifying and then testing partners that may have been exposed to an infectious agent allows additional information about the partners to be collected. While the primary purpose of contact tracing is disease control via partner notification and partner services, it is also a form of data collection that is rarely utilized. Such approaches are used most commonly for syphilis and HIV/AIDS, but also for other STIs such as gonorrhea and chlamydia [12], as well as routinely for tuberculosis and infectious disease outbreaks. Contact tracing has also been applied in many recent epidemics, including COVID-19 [1, 4, 5, 27]. In *positive contact tracing*, we follow all ties from infected nodes, but ties from uninfected nodes are not followed.

While the exact process varies with the public health situation, we consider the following stylized *biased seed link tracing process*:

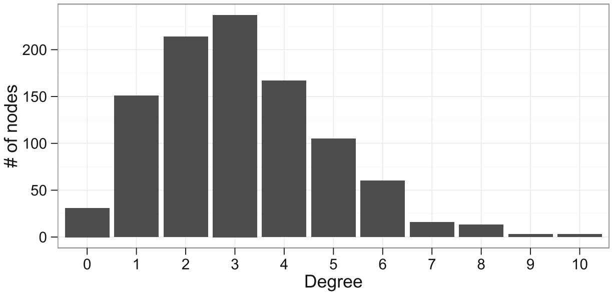


**Fig. 4** Degree distribution of the networked population

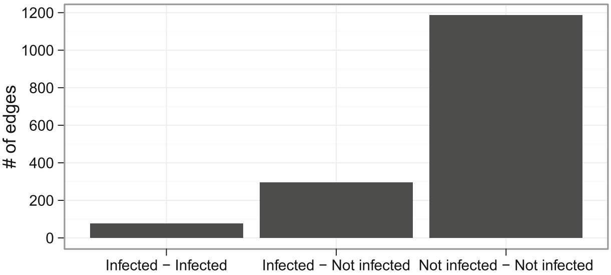


**Fig. 5** Mixing statistics: Counts of the numbers of ties by the infection status of the incident nodes for the networked population

1. Select $s_{-\text{inf}}$ seed subjects at random from among the non-infected population and observe them.
2. Select $s_{\text{inf}}$ seeds subjects at random from among the infected population and observe them.
3. Choose the next infected seed at random from among the infected seeds.
4. Observe all ties from the selected subject, and the infection status of these subjects.
5. For all infected neighbors of the selected subject, go to Step 4.
6. If all the seeds have not been chain sampled as part of Step 3, go to Step 3

We simulated a networked population of $n = 1000$ people from the simple homophily model of Sect. 2.1 with natural parameters of $\eta = (-5.8, .7, -1.95)$. The number of infected nodes was fixed at 150. The generated network had a mean degree of 3.1, and its degree distribution is displayed in Fig. 4. Figure 5 displays the number of ties categorized by the infection status of their incident nodes. For example, there were 296 infected to non-infected ties. This indicates moderate homophily on tie status (via a permutation test, not shown here).

Starting with $s_{\text{inf}} = 40$ infected seeds, we simulated 100 positive link tracing samples for each of $s_{-\text{inf}} = (0, 45, 90, 135, 180, 225)$. Figure 6 displays a histogram of the sizes of the samples when there are no non-infected seeds (i.e., $s_{-\text{inf}} = 0$).

To provide a comparison for our method we considered two estimators that could be utilized. Neither of them uses a model for the networked population but is motivated by approximations to the sampling design. The first treats the sample as a simple random sample:

$$\text{Naive} = n\frac{n_{\text{inf}}}{n_{\text{inf}} + n_{-\text{inf}}},$$

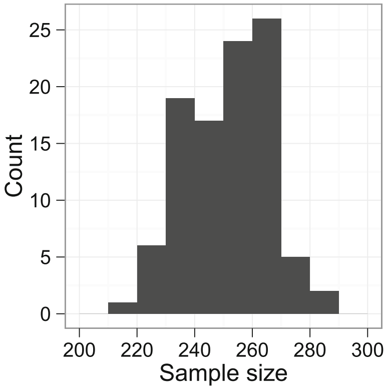


**Fig. 6** Sizes of the contact-traced samples based on 40 seed subjects ($s_{\text{inf}} = 40$, $s_{-\text{inf}} = 0$)

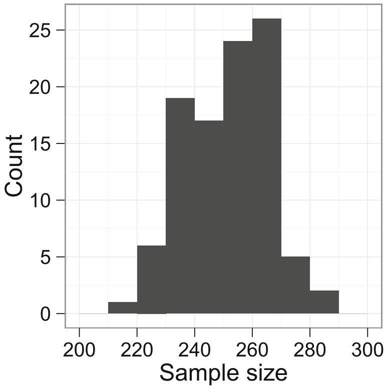


**Fig. 7** Estimates via contact tracing with $s_{\text{inf}} = 40$ infected seeds and varying numbers of non-infected seeds

The second adjusts for the sampling of the seeds:

$$\text{Naive (seed adj.)} = (n - s_{\text{inf}} - s_{-\text{inf}}) \frac{n_{\text{inf}} - s_{\text{inf}}}{n_{\text{inf}} - s_{\text{inf}} + n_{-\text{inf}} - s_{-\text{inf}}} + s_{\text{inf}}.$$

Our approach is to fit an ERNM to the contact tracing data. In this situation the contact tracing sampling design is clearly informative. For comparison, we compute two estimates of the model. The first takes into account the informativeness of the contact tracing design (MNAR) and the other assumes it is ignorable (MAR). These are based on the likelihoods (2) and (5), respectively, and the algorithm in Sect. 4.

Figure 7 shows the results for each of the estimators over the samples. The median of the MNAR estimator is centered around the true value of 150 in all sampling scenarios, while the MAR estimator performs poorly with all infected seeds ($s_{-\text{inf}} = 0$) and increasingly well as the number of non-infected seeds increases to $s_{-\text{inf}} = 225$. This is somewhat expected as the proportion of infected in the seeds approximately matches that of the population when $s_{-\text{inf}} = 225$. The two naive estimators are significantly biased across all samples. This is especially true for the sample mean which is biased both by the seed selection and by the link-tracing design. The adjusted sample mean corrects somewhat for the seed bias but does not represent the link-tracing.

This application illustrates the advantage of the model-based approach over the *ad hoc* estimators. By representing the structure of the networked population, the model-based approach can leverage the information in the data more efficiently.

## 6 Discussion

In this paper we have given a concise and systematic statistical framework for dealing with partially observed network data when some knowledge is available on the sampling design. The framework includes, but is not restricted to, ignorable sampling designs. We have also shown that likelihood-based inference is practical under partial observation for ERNMs, and that the likelihood framework naturally accommodates standard network sampling designs.

We developed and implemented algorithms to compute Monte Carlo approximations to the likelihood, and showed how these can be used in practice. Three important special cases of these designs were demonstrated in Sect. 5. In Sect. 5.1 we consider a missingness process which randomly selected tie variables and nodal attributes to be missing. In Sect. 5.2 we consider sampling designs where the seeds are not chosen at random from the population, but are some form of convenience sample chosen by an, at least partially, unknown process. We show that there are cases where we can compute the observed data likelihood without knowing the specific mechanism of seed selection. For example, it shows that this is true when the mechanism for selecting seeds from the uninfected population is unknown, as long as the seeds from the infected population are chosen by a known design. In Sect. 5.3 we consider non-ignorable sampling in the context of contact tracing data, a case of vital importance to public health. It considers a sampling design where all ties from infected nodes are followed/sampled, but ties from uninfected nodes are not followed/sampled. At present, this is the first statistically defensible approach to inference in this form of data. The example presented here shows that the MLE estimation task is robust, in that it can be applied successfully to moderately large networks (1000 nodes), with significant missingness (greater than 70% of nodes unobserved).

The methods developed in this paper have been implemented in an open-source R package [2, 21]. It was used to do the simulation studies and analyze the case-studies, and is available for general use.

**Acknowledgements** We are grateful for support from the National Science Foundation BIGDATA: Applications program, grant NSF IIS-1546259, grants MMS-0851555 and SES-1357619, and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, for grant R21HD075714, population research infrastructure grants P2C-HD041041 and P2C-HD041022 and training grant T32-HD007545.

## Declarations

**Conflict of interest** The authors declare no competing interests.

## Appendix: Algorithmic and computational details

### A.1 Alternate MLE formulation

While the algorithm outlined in Sect. 4 works well, there are some situations where an alternate formulation using Eq. (5) may be useful. First let us consider the case where $\theta = \theta_0$, then the likelihood is

$$\ell(\eta) - \ell(\eta_0) = \log\left(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(Z)} \mid z_{obs})\right) - \log\left(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(Z)})\right) \qquad (1)$$
$$+ \log\left(\frac{E_{\eta}(p(W = w \mid Z, \theta) \mid Z_{obs} = z_{obs})}{E_{\eta_0}(p(W = w \mid Z, \theta) \mid Z_{obs} = z_{obs})}\right)$$

The first expectation, and the expectation in the denominator of the third term, can be calculated using an MCMC sample from $p(z \mid z_{obs}, \eta_0)$. The second can be approximated with an MCMC sample from $p(z \mid \eta_0)$. The numerator of the third term can be approximated by importance sampling.

$$E_\eta(p(W = w \mid Z, \theta) \mid Z_{obs} = z_{obs}) \approx \frac{1}{M} \sum_{k=1}^{M} p(w \mid z^{(k)}, \theta)\omega^{(k)}$$

where $z^{(k)} \sim p(z \mid z_{obs}, \eta_0)$ and

$$\omega^{(k)} = \frac{e^{(\eta-\eta_0)\cdot g(z^{(k)})}}{\sum_{j=1}^{M} e^{(\eta-\eta_0)\cdot g(z^{(j)})}}$$

If the sampling process is ignorable, then the third term drops out of the likelihood ratio. The first and second derivatives of the likelihood are useful in the maximization process. For notational convenience, let $\Delta_i(t) = g_i(t) - E(g_i(Z))$.

$$\begin{aligned}\left[\frac{\delta \ell}{\delta \eta}\right]_i &= \frac{\delta}{\delta \eta_i} \log\left(\sum_{z_{miss}} p(W = w \mid Z = z) p(Z_{miss} = z_{miss} \mid \eta, Z_{obs} = z_{obs}) p(Z_{obs} = z_{obs} \mid \eta)\right)\\ &= \frac{\sum_{z_{miss}} p(W = w \mid Z = z)\Delta_i(t) p(Z_{miss} = z_{miss} \mid \eta, Z_{obs} = z_{obs}) p(Z_{obs} = z_{obs} \mid \eta)}{\sum_{z_{miss}} p(W = w \mid Z = z) p(Z_{miss} = z_{miss} \mid \eta, Z_{obs} = z_{obs}) p(Z_{obs} = z_{obs} \mid \eta)}\\ &= \frac{E(p(W = w \mid Z)\Delta_i(Z) \mid Z_{obs} = z_{obs})}{E(p(W = w \mid Z) \mid Z_{obs} = z_{obs})}\end{aligned}$$

$$\begin{aligned}&\left[\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j}\right]_{ij}\\ &= \frac{\delta}{\delta \eta_j} \frac{\sum_{z_{miss}} p(W = w \mid Z = z)\Delta_i(t) p(Z_{miss} = z_{miss} \mid \eta, Z_{obs} = z_{obs}) p(Z_{obs} = z_{obs} \mid \eta)}{\sum_{z_{miss}} p(W = w \mid Z = z) p(Z_{miss} = z_{miss} \mid \eta, Z_{obs} = z_{obs}) p(Z_{obs} = z_{obs} \mid \eta)}\\ &= -\mathrm{cov}(g_i(Z), h_j(Z)) + \frac{E(p(W = w \mid Z)\Delta_i(Z)\Delta_j(Z) \mid Z_{obs} = z_{obs})}{E(p(W = w \mid Z) \mid Z_{obs} = z_{obs})}\\ &\quad - \frac{E(p(W = w \mid Z)\Delta_i(Z) \mid Z_{obs} = z_{obs}) E(p(W = w \mid Z)\Delta_j(Z) \mid Z_{obs} = z_{obs})}{E(p(W = w \mid Z) \mid Z_{obs} = z_{obs})^2}\end{aligned} \tag{2}$$

If the missingness process is ignorable, these equations simplify to

$$\begin{aligned}\left[\frac{\delta \ell}{\delta \eta}\right]_i &= E(\Delta_i(Z) \mid Z_{obs} = z_{obs})\\ \left[\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j}\right]_{ij} &= -\mathrm{cov}(g_i(Z), g_j(Z)) + \mathrm{cov}(g_i(Z), g_j(Z) \mid Z_{obs} = z_{obs})\end{aligned}$$

If we fix $\eta$, then the observed data likelihood of $\theta$

$$\begin{aligned}L(\theta \mid z_{obs}, w, \eta) &\propto p(z_{obs} \mid \eta) E(p(W = w \mid Z, \theta) \mid Z_{obs} = z_{obs})\\ &= E(p(W = w \mid Z, \theta) \mid Z_{obs} = z_{obs}, \eta)\end{aligned}$$

can be maximized to find the MLE of $\theta$.

This motivates the following algorithm for maximizing the observed data likelihood.

1. Let $l = 0$ and choose initial parameter values $\eta^{(0)}$, $\theta_0$.
2. Use MCMC to generate $M_{\text{miss}}$ samples, $z_{miss}^{(k)}$, from $p(z_{miss} \mid \eta^l, z_{obs})$.
3. Use MCMC to generate $M_{\text{full}}$ samples, $z^{(k)}$, from $p(z \mid \eta^l)$.

4. Set $\theta^{l+1} = \text{argmax}_\theta(E(p(w \mid Z, \theta) \mid Z_{obs} = z_{obs}, \eta))$, with samples from Step 2 used to approximate the expectation.
5. Using the samples from Steps 2 and 3 to approximate the relevant expectations, find $\eta^{l+1}$ maximizing Eq. (1) subject to the convergence condition $\|\eta^{l+1} - \eta^l\| < \epsilon$.
6. If the likelihood has not converged, set $l = l + 1$ and go to Step 2.
7. Set $l = l + 1$, and go to Step 2.

The disadvantage of this method is that if the networks generated by the MNAR process are very different from those generated assuming MAR, the estimates of the last expectation in Eq. (1) can have high variance. The benefit of using this method is that the sampling probability, $p(W = w \mid Z = z, \theta)$, only needs to be calculated for networks included in the sample, and not at every MCMC step as is required by the algorithm in Sect. 4, so if the sampling probability is computationally expensive to calculate, this method can be significantly faster than the one outlined in Sect. 4

### A.2 Estimating network statistics

We can use MCMC samples from $p(z_{miss} \mid z_{obs}, \eta)$ to estimate the (complete) network statistics of the sampled network. Suppose that we have used MCMC to draw $M$ samples $z^{(k)}_{miss}$ from the distribution $p(z_{miss} \mid z_{obs}, \eta)$, and $z^{(k)} = (z_{obs}, z^{(k)}_{miss})$. Then we can estimate the expectation of a set of network statistics $g$ as

$$E(g(Z) \mid z_{obs}, \eta) \approx \frac{1}{M} \sum_{k=1}^{M} g(z^{(k)}).$$

However, this equation ignores the possible bias introduced by our sampling process $w$. The distribution that we should be sampling from is the full conditional distribution of $z_{miss}$,

$$\begin{aligned} p(Z_{miss} &= z_{miss} \mid Z_{ob} = z_{obs}, W = w, \eta) \\ &\propto p(Z_{miss} = z_{miss} \mid Z_{obs} = z_{obs}, \eta) p(W = w \mid Z = z, \theta). \end{aligned}$$

We then use importance sampling to estimate the relevant quantity

$$E(g(Z) \mid z_{obs}, w, \eta, \theta) \approx \frac{\sum_{k=1}^{M} g(z^{(k)}) p(W = w \mid Z = z^{(k)}, \theta)}{\sum_{k=1}^{M} p(W = w \mid Z = z^{(k)}, \theta)}.$$

## References

1. Donnelly, C.A., Ghani, A.C., Leung, G.M., Anderson, R.M., et al.: Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet **361**(9371), 1761–1766 (2003). https://doi.org/10.1016/S0140-6736(03)13410-1
2. Fellows, I.E.: ernm: exponential-family random network models (2014–2022). R package version 1.1. https://github.com/fellstat/ernm
3. Fellows, I.E., Handcock, M.S.: Exponential-family random network models (2012). arXiv:1208.0121 [stat.ME]
4. Fenner, F., Henderson, D.A., Arita, I., Jezek, Z., Ladnyi, I.D.: Smallpox and Its Eradication vol. 6. World Health Organization, Geneva (1988). https://www.aphl.org/programs/preparedness/Smallpox/pdf/9241561106.pdf
5. Ferguson, N.M., Donnelly, C.A., Anderson, R.M.: Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature **413**(6855), 542–548 (2001). https://doi.org/10.1038/35097116

6. Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dorner, L., Parker, M., Bonsall, D., Fraser, C.: Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. Science (2020). https://doi.org/10.1126/science.abb6936
7. Frank, O., Strauss, D.: Markov graphs. J. Am. Stat. Assoc. **81**(395), 832–842 (1986). https://doi.org/10.2307/2289017
8. Gile, K.J.: Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. J. Am. Stat. Assoc. **106**(493), 135–146 (2011). https://doi.org/10.1198/jasa.2011.ap09475
9. Gile, K.J., Handcock, M.S.: Network model-assisted inference from respondent-driven sampling data. J. R. Stat. Soc. A. Stat. Soc. **178**, 619–639 (2015). https://doi.org/10.1111/rssa.12091
10. Gile, K.J., Handcock, M.S.: Respondent-driven sampling: an assessment of current methodology. Sociol. Methodol. **40**, 285–327 (2010). https://doi.org/10.1111/j.1467-9531.2010.01223.x
11. Gile, K.J., Handcock, M.S.: Analysis of networks with missing data with application to the national longitudinal study of adolescent health. J. R. Stat. Soc. Ser. C (Appl. Stat.) **66**, 501–519 (2016). https://doi.org/10.1111/rssc.12184
12. Golden, M.R., Hogben, M., Potterat, J.J., Handsfield, H.H.: HIV partner notification in the United States: a national survey of program coverage and outcomes. Sex Transm. Dis. **31**(12), 709–712 (2004). http://www.jstor.org/stable/44966741
13. Handcock, M.S., Gile, K.J.: Modeling networks from sampled data. Ann. Appl. Stat. **272**(2), 383–426 (2010). https://doi.org/10.1214/08-AOAS221
14. Handcock, M.S., Gile, K.J., Mar, C.M.: Estimating hidden population size using respondent-driven sampling data. Electron. J. Stat. **8**(1), 1491–1521 (2014). https://doi.org/10.1214/14-EJS923
15. Handcock, M.S., Gile, K.J., Mar, C.M.: Estimating the size of populations at high risk for HIV using respondent-driven sampling data. Biometrics **71**(1), 258–266 (2015). https://doi.org/10.1111/biom.12255
16. Harris, K.M., Florey, F., Tabor, J., Bearman, P.S., Jones, J., Udry, J.R.: The national longitudinal of adolescent health: Research design [WWW document]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill (2003). https://doi.org/10.17615/C6TW87
17. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. Am. Stat. Assoc. **97**(460), 1090–1098 (2002). https://doi.org/10.1198/016214502388618906
18. Hunter, D.R., Handcock, M.S.: Inference in curved exponential family models for networks. J. Comput. Graph. Stat. (2006). https://doi.org/10.1198/106186006X133069
19. Klinkenberg, D., Fraser, C., Heesterbeek, H.: The effectiveness of contact tracing in emerging epidemics. PLoS One **1**(1), 12 (2006). https://doi.org/10.1371/journal.pone.0000012
20. Potterat, J.J., Spencer, N.E., Woodhouse, D.E., Muth, J.B.: Partner notification in the control of human immunodeficiency virus infection. Am. J. Public Health **79**(7), 874–876 (1989). https://doi.org/10.2105/ajph.79.7.874
21. R Core Team: R.: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2020). https://www.R-project.org/
22. Rubin, D.: Inference and missing data. Biometrika **63**, 581–592 (1976). https://doi.org/10.2307/2335739
23. Sampson, S.F.: Crisis in a cloister. PhD in Sociology, Cornell University (1969). https://doi.org/10.1111/j.1467-9531.2010.01223.x
24. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)
25. Thompson, S.K.: Adaptive and network sampling for inference and interventions in changing populations. J. Surv. Stat. Methodol. **5**(1), 1–21 (2017). https://doi.org/10.1093/jssam/smw035. https://academic.oup.com/jssam/article-pdf/5/1/1/29005492/smw035.pdf
26. Udry, J.R.: The national longitudinal of adolescent health: (add health), waves I and II, 1994–1996; wave III, 2001–2002 [machine-readable data file and documentation]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill (2003). https://www.disc.wisc.edu/codebooks/qg-067-002.pdf
27. Yuan, H.-Y., Blakemore, C.: The impact of contact tracing and testing on controlling Covid-19 outbreak without lockdown in Hong Kong: an observational study. The Lancet Regional Health Western Pacific (2022). https://doi.org/10.1016/j.lanwpc.2021.100374