Modeling Social Networks with Sampled Data [1]

Technical Report no. 523
Department of Statistics
University of Washington

Mark S. Handcock
Krista Gile
University of Washington, Seattle

Revised December 5, 2007

[1] Mark S. Handcock is Professor of Statistics, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: handcock@stat.washington.edu; Web: www.stat.washington.edu/handcock. Krista Gile is in the Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4332. E-mail: kgile@stat.washington.edu; Web: http://www.stat.washington.edu/kgile/; This research was partially supported by Grant DA012831 from NIDA and Grant HD041877 from NICHD.

Abstract

Network models are widely used to represent relational information among interacting units and the structural implications of these relations. Recently, social network studies have focused a great deal of attention on random graph models of networks whose nodes represent individual social actors and whose edges represent a specified relationship between the actors.

Most inference for social network models assumes that the presence or absence of all possible links is observed, that the information is completely reliable, and that there are no measurement (e.g. recording) errors. This is clearly not true in practice, as much network data is collected though sample surveys. In addition even if a census of a population is attempted, individuals and links between individuals are missed (i.e., do not appear in the recorded data).

In this paper we develop the conceptual and computational theory for inference based on sampled network information. We first review forms of network sampling designs used in practice. We consider inference from the likelihood framework, and develop a typology of network data that reflects their treatment within this frame. We then develop inference for social network models based on information from adaptive network mechanisms.

We motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on a collaboration network.

1 Introduction

Networks are a useful device to represent "relational data", that is, data with properties beyond the attributes of the individuals (nodes) involved. Relational data arise in many fields and network models are a natural approach to representing the patterns of the relations between nodes. Networks can be used to describe such diverse ideas as the behavior of epidemics, the interconnectedness of corporate boards, and networks of genetic regulatory interactions. In social network applications, the nodes in a graph typically represent individuals, and the ties (edges) represent a specified relationship between individuals. Nodes can also be used to represent larger social units (groups, families, organizations), objects (airports, servers, locations), or abstract entities (concepts, texts, tasks, random variables). We consider here stochastic models for such graphs. These models attempt to represent the stochastic mechanisms that produce relational ties, and the complex dependencies thus induced.

Social network data typically consist of a set of n actors and a relational tie random variable, Y_{ij}, measured on each possible ordered pair of actors, $\{i, j\}$, $i, j = 1, \ldots, n., i \neq j$. In the most simple cases, Y_{ij} is a dichotomous variable, indicating the presence or absence of some relation of interest, such as friendship, collaboration, transmission of information or disease, etc. The data are often represented by an $n \times n$ sociomatrix Y, with diagonal elements, representing self-ties, treated as structural zeros. In the case of binary relations, the data can also be thought of as a graph in which the nodes are actors and the edge set is $\{(i, j) : Y_{ij} = 1\}$. For many networks the relations are undirected in the sense that $\{Y_{ij} = Y_{ji}, i, j = 1, \ldots, n\}$.

For large or hard to find populations of actors it is difficult to obtain information on all actors and all relational ties. As a result various survey sampling strategies and methods are applied. Some of these methods make use of network information revealed by earlier stages of sampling to guide later sampling. These adaptive designs allow for more efficient sampling than conventional sampling designs. We consider such designs in Section 2.

In this paper we mainly consider the network over the set of actors to be the realization of a stochastic process. We seek to model that process. An alternative is to view the network as a fixed structure about which we wish to make inference based on partial observation.

In this paper we develop a theoretical framework for inference from network data that are partially-observed due to sampling. This work extends the fundamental work of Thompson and Frank (2000). For purposes of presentation, we focus on the relational data itself and suppress reference to covariates of the nodes. This more general situation is dealt with in Handcock and Gile (2007).

In Section 2 we present a conceptual framework for network sampling. We extend this framework in Section 3 to focus on inference from sampled network data. We consider both design-based and model-based inference. Section 4 presents the rich Exponential Family Random Graph Model (ERGM) family of models that has been applied to complete network data. Section 5 presents a study of the effect of sampling from a known complete network of law firm collaborations. Finally, in Section 6, we discuss the overall ramifications for the modeling of social networks with sampled data and note some extensions.

2 Network Sampling Design

In this section we consider the conceptual and computational theory of network sampling.

There is a substantial literature on network sampling designs. Our development here follows Thompson and Seber (1996) and Thompson and Frank (2000). Let $\mathcal{Y}$ denote the set of possible networks on the n actors. Note that in most network samples, the unit of sampling is the actor or node, while the unit of analysis is typically the dyad. Let D be the $n \times n$ random binary matrix indicating if the corresponding element of Y was sampled or not. The value of the i, j^{th} element is 0 if the i, j^{th} ordered pair was not sampled and 1 if the element was sampled. We shall refer to D as the *network design mechanism.* We shall refer to realizations of D, from some sample space $\mathcal{D}$, as the *design matrix* and the probability distribution of D as the *design mechanism.* The design mechanism is often related to the structure of the graph and a parameter $\psi \in \Psi$, so we posit a model for it. Specifically, let $pr(D = d|Y = y; \psi)$ denote the probability of the design mechanism selecting sample d given a network y and parameter ψ.

Under many sampling designs the set of sampled dyads is determined by the set of sampled nodes. Let S represent a binary random $n-$vector indicating a subset of the nodes, where the i^{th} element is 1 if the i^{th} node is part of the set, and is 0 otherwise. We often consider situations where D is determined by some S which is itself a result of a design mechanism denoted by $pr(S|Y, \psi)$. For example, consider an undirected network where the set of observed dyads are those that are incident on at least one of the sampled nodes. In this case $D = S1^T + 1S^T - SS^T$, where 1 is the binary $n-$vector of 1s. A primary example of this is where people are sampled and surveyed to determine all their edges.

We introduce further notation to allow us to refer to the observed and unobserved portions of the relational structures. Denote the observed part of the complete graph Y by $Y_{obs} = \{Y_{ij} : D_{ij} = 1\}$ and the unobserved part by $Y_{mis} = \{Y_{ij} : D_{ij} = 0\}$. The full *observed data* is then $\{Y_{obs}, D\}$, in contrast to the *complete data:* $\{Y_{obs}, Y_{mis}, D\}$. We will write the complete graph $Y = \{Y_{obs}, Y_{mis}\}$. In addition, if we make the convention that a number plus or multiplied by an undefined number is the number, we have $Y = Y_{obs} + Y_{mis}$. For a given network $y \in \mathcal{Y}$, denote the corresponding data as $\{y_{obs}, d\}$ and the other elements by their lower-case versions $y = y_{obs} + y_{mis}$.

A design mechanism is *conventional* if it does not use information collected during the survey to direct subsequent sampling of individuals (e.g., network census and ego-centric designs). Specifically, a design is conventional if $pr(D = d|Y = y; \psi) = pr(D = d|\psi)\ \forall y \in \mathcal{Y}$. A simple example of a conventional design mechanism for networks is simple random sampling of a subset of the actors, followed by complete observation of the dyads originating from those actors. A complete census of the network is another. More complex examples include designs using probability sampling of pairs and auxiliary variables. Alternatively, we call a design mechanism *adaptive* if it uses information collected during the survey to direct subsequent sampling, but the design mechanism depends only on the observed data. Specifically, a design is adaptive if: $pr(D = d|Y = y; \psi) = pr(D = d|Y_{obs} = y_{obs}, \psi)\ \forall y \in \mathcal{Y}$. Conventional designs can be considered to be special cases of adaptive designs.

Note that the mechanism for adaptive designs satisfies

$$pr(D = d|Y_{obs}, Y_{mis}, \psi) = pr(D = d|Y_{obs}, \psi),$$

a condition called "*missing at random*" by Rubin (1976) in the context of missing data. Note that this is a bit of a misnomer – it does not say that the propensity to be observed is unrelated to the unobserved portions of the network, but that this relationship can be explained by the data that are observed. The observed part of the data are often vital to this equality. Hence adaptive designs are essentially those for which the unobserved dyads are missing at random.

2.1 Some Adaptive designs for Undirected Networks

We now consider several examples of adaptive designs for undirected networks.

2.1.1 Example: Ego-Centric design

Consider a simple *ego-centric design*:

1. Select individuals at random, each with probability ψ.

2. Observe all dyads involving the selected individuals (i.e., dyads with at least one of the selected individuals as one of the pair of actors).

The sampling mechanism can be determined for this design. First note that

$$pr(D_{ij} = 1|Y, \psi) = 1 - (1 - \psi)^2 \ \ \forall i \neq j$$

This, however, does not give the joint distribution of D. Denote by $[y]$ the vector-valued function that is 1 if the corresponding element of y is logically true, and 0 otherwise. Let $y \cdot x$ be the elementwise product of y and x. Let S be the binary $n-$vector where 1 and 0 indicate that the corresponding individual has been selected, or not, respectively. Within this design, S is determined by the design matrix (i.e. $S = [D1 = (n-1)1]$). Then $pr(S = s|Y, \psi) = \psi^{1^T s}(1-\psi)^{n-1^T s} \ \ s \in \{0,1\}^n$. If the ith element of S is 1 then all elements in the ith row and column of D are 1. $D_{ij} = 0$ if and only if both the ith and jth elements of S are both 0. Hence the probability distribution of D is:

$$pr(D = d|Y, \psi) = \psi^{1^T s}(1-\psi)^{n-1^T s}$$

for

$$d = 1s^T + s1^T - ss^T \ \ s \in \{0,1\}^n$$

Note that the distribution does not depend on Y.

2.1.2 Example: One-wave link-tracing design

We refer to any sample in which subsequent nodes are enrolled based on their observed relations with other sampled nodes and a *link-tracing* design. Consider the one-wave link-tracing design specified as follows:

1. Select individuals at random, each with probability ψ.

2. Observe all dyads involving the selected individuals.

3. Identify all individuals reported to have at least one relation with the initial sample, and select them with probability 1.

4. Observe all dyads involving the newly selected individuals.

Let S_0 denote the indicator vector for the initial sample and S_1 the indicator for the added individuals not in the initial sample. Then the whole sample of individuals is $S = S_0 + S_1$. As in the undirected ego-centric design, the design matrix is given by $D = 1S^T + S1^T - SS^T$. Note that $S_1 = [YS_0 \cdot (1 - S_0)]$ is derivable from S_0 and Y. Hence

$$pr(D = d|Y, \psi) = \sum_{s_0:\ s_0+[Ys_0\cdot(1-s_0)]=s} \psi^{1^T s_0}(1-\psi)^{n-1^T s_0}$$

for

$$d = 1s^T + s1^T - ss^T \quad s \in \{0,1\}^n$$

2.1.3 Example: Multi-wave link-tracing design

Consider a *multi-wave link-tracing design* in which the complete set of partners of the kth wave are enrolled, i.e. the link-tracing process described above is carried out k times.

Let S_0 denote the indicator for the initial sample, S_1 the indicator for the added individuals in the first wave not in the initial sample, $\ldots$, S_k the indicator for the added individuals in wave k not in the prior samples. Then the whole sample of individuals is $S = S_0 + S_1 + \ldots + S_k$. As in the ego-centric design the design matrix is given by $1S^T + S1^T - SS^T$. Note that $S_m = \left[YS_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right]$,
$m = 1, \ldots, k$ is derivable from S_0 and Y . Then:

$$pr(D = d|Y, \psi) = \sum_{s_0:\ s_0+s_1+\ldots+s_k=s} \psi^{1^T s_0}(1-\psi)^{n-1^T s_0}$$

for $d = 1s^T + s1^T - ss^T \quad s \in \{0,1\}^n$. Here $S_m = \left[YS_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right] = \left[Y_{obs}S_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right]$, $m = 1, \ldots, k$ so that the individuals selected in the successive waves only depend on the observed part of the graph, and not on the unobserved portions of the graph. Clearly, this is also true for one-wave link-tracing as a simple case of $k-$wave link-tracing. If k is fixed in advance this is called *$k-$wave link-tracing*. Note that it may be possible that $S_m = S_{m-1}$ for some $m < k$, so that subsequent waves do not increase the sample size (i.e., $S_k = S_{m-1}$). A variant of the $k-$wave link-tracing design is the *total-wave link-tracing* design that sets $k = \infty$.

2.2 Some Adaptive Designs for Directed Networks

We can also consider variants of these adaptive designs for directed networks.

2.2.1 Example: Ego-Centric design

Consider a simple *ego-centric design*:

1. Select individuals at random, each with probability ψ.

2. Observe all out-arcs and non-arcs from the selected individuals

As before, the sampling mechanism can be determined for this design. Since a directed dyad is observed only if its tail node is sampled,

$$pr(D_{ij} = 1|Y, \psi) = \psi \ \ \forall i \neq j$$

and $D = S_0 1^T$. Hence the probability distribution of D is:

$$pr(D = d|Y, \psi) = \psi^{1^T s}(1 - \psi)^{n-1^T s}$$

for $d = s1^T \ \ s \in \{0, 1\}^n$ and the distribution does not depend on Y.

2.2.2 Example: One-wave link-tracing design

Consider a one-wave link-tracing design on a directed network specified as follows:

1. Select individuals at random, each with probability ψ.
2. Observe all out-arcs and non-arcs from the selected individuals
3. Identify all individuals receiving an arc from a member of the initial sample, and select them with probability 1.
4. Observe all out-arcs and non-arcs from the newly selected individuals

Let S_0 denote the indicator vector for the initial sample and S_1 the indicator for the added individuals not in the initial sample. Then the whole sample of individuals is $S = S_0 + S_1$. As in the ego-centric design the design matrix is given by $D = S1^T$ and

$$pr(D = d|Y, \psi) = \sum_{s_0\colon\, s_0+[Ys_0\cdot(1-s_0)>0]=s} \psi^{1^T s_0}(1 - \psi)^{n-1^T s_0}$$

for $d = s1^T \ \ s \in \{0, 1\}^n$.

2.2.3 Example: Multi-wave link-tracing design

Consider a directed version of the multi-wave link-tracing design in which the complete set of out-partners of the kth wave are enrolled. The whole sample of individuals is $S = S_0 + S_1 + ... + S_k$. and $S_m = \left[YS_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right]$, $\ m = 1, ..., k$ is derivable from S_0 and Y . Then:

$$pr(D = d|Y, \psi) = \sum_{s_0\colon\, s_0+s_1+...+s_k=s} \psi^{1^T s_0}(1 - \psi)^{n-1^T s_0}$$

for $d = s1^T \ \ s \in \{0, 1\}^n$, where we note that $S_m = \left[YS_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right] = \left[Y_{obs}S_{m-1} \cdot (1 - \sum_{t=0}^{m-1} S_{t-1}) > 0\right]$, $m = 1, ..., k$ so that the individuals selected in the successive waves of depend only on the observed part of the graph, and not on the unobserved portions of the graph. The total-wave link-tracing design is defined by $k = \infty$.

3 Inferential Frameworks

In this section we consider two frameworks for inference based on sampled data. In the *design-based* framework y represents the population and interest focuses on characterizing y based on partial observation. Under the *model-based* framework Y is stochastic and is a realization from a stochastic process depending on a parameter η. Here interest focuses on η which characterizes the mechanism that produced the complete network Y. The model may also be used to guide design-based inference (Särndal et al., 1992). We consider the design and model-based frameworks in turn.

3.1 Design-based inference for the Network

In the design-based frame, the unobserved data values, or some functions thereof, are analogous to the parameters of interest in likelihood inference. The population of data values is treated as fixed, and all uncertainty in the estimates is due to the sampling mechanism, which is typically assumed to be fully known (not just up to the parameter ψ.).

Inference typically focuses on identifying design-unbiased estimators for quantities of interest measured on the complete network. In an undirected network analysis setting, for example, we can consider estimating $\tau = \sum_{i<j} y_{ij}$, the number of edges in the network. Note that y is a partially-observed matrix of constants in this setting. Then $\hat{\tau}$ is design-unbiased for τ if:

$$\mathbb{E}_D[\hat{\tau}|\psi, y] = \tau,$$

where the expectation is taken over realizations of the sampling process. Specifically:

$$\mathbb{E}_D[\hat{\tau}(Y_{obs}, D)|\psi, y] = \sum_{d \in \mathcal{D}} \hat{\tau}(y_{obs}(d), d) pr(D = d|\psi, y),$$

where $\hat{\tau}(y_{obs}(d), d)$ is the estimator expressed as a function of the observed network information. Similarly, the variance of the estimator is computed with respect to the variation induced by the sampling procedure:

$$\mathbb{V}_D[\hat{\tau}(Y_{obs}, D)|\psi, y] = \sum_{d \in \mathcal{D}} (\hat{\tau}(y_{obs}(d), d) - \tau)^2 pr(D = d|\psi, y),$$

The Horvitz-Thompson estimator is a classic tool of design-based inference, and is based on inverse-probability weighting the sample. In our example, it is:

$$\hat{\tau}(Y_{obs}, D) = \sum_{ij:D_{ij}=1} \frac{y_{ij}}{\pi_{ij}},$$

where the *dyadic sampling probability* $\pi_{ij} = pr(D_{ij} = 1|\psi, y)$ is the probability of observing dyad (i, j).

Consider an estimator of τ based on relations observed through the ego-centric design of Section 2.1.1. Then:

$$\pi_{ij} = 1 - (1 - \psi)^2 \ \ \forall \ \ i, j.$$

The classic Horvitz-Thompson estimator $\hat{\tau}$ of τ then weights each observation by the inverse of its sampling probability:

$$\hat{\tau} = \sum_{ij:D_{ij}=1} \frac{y_{ij}}{\pi_{ij}} \quad = \quad \frac{1}{1-(1-\psi)^2} \sum_{ij:D_{ij}=1} y_{ij}.$$

Then

$$\mathbb{V}(\hat{\tau}) = \sum_{i<j} \sum_{k<l} \left\{ [1-(1-\psi)^2]^{-2} \pi_{ij,kl} - 1 \right\} y_{ij} y_{kl},$$

where $\pi_{ij,kl} = pr(\exists\ u, v \ :\ S_{0u} = S_{0v} = 1 \implies D_{ij} = D_{kl} = 1)$ or:

$$\pi_{ij,kl} = \begin{cases} \pi_{ij} & i = k,\ j = l \\ \pi_{ij}\pi_{kl} & i \notin \{k, l\} \text{ and } j \notin \{k, l\} \\ \psi^3 - 3\psi^2 & \text{otherwise} \end{cases}$$

Among the many available estimators for the variance of the Horvitz-Thompson estimator is the Horvitz-Thompson variance estimator:

$$\hat{\mathbb{V}}(\hat{\tau}) = \sum_{ij:D_{ij}=1} \sum_{kl:D_{kl}=1} \frac{1}{\pi_{ij,kl}} \left\{ [1-(1-\psi)^2]^{-2} \pi_{ij,kl} - 1 \right\} y_{ij} y_{kl}.$$

Note the importance of the unit sampling probabilities in these estimators. This is a hallmark of design-based inference: inference relies on full knowledge of the sampling procedure in order to make unbiased inference without making assumptions about the distribution of the unobserved data. This typically requires knowledge of the sampling probability of each unit in the sample. This procedure is complicated in the network context, in that we require the sampling probabilities of the units of analysis, dyads, which are different from the units of sampling, nodes. In fact, for even single-wave link-tracing samples, the dyadic sampling probabilities are not observable.

To see this, define the *nodal neighborhood of a dyad* (i, j), $N(i, j)$, where $k \in N(i, j) \iff \{S_{0k} = 1 \implies D_{ij} = 1\}$. Then $\pi_{ij} = pr(\exists\ k\ :\ S_{0k} = 1, k \in N(i, j))$.

For the one-wave link-tracing design of Section 2.1.2, $N(i, j) = \{k\}\ :\ y_{ik} = 1 or y_{jk} = 1 or k \in \{i, j\}$. Then if the initial sample S_0 is drawn according to the design in 2.1.2, $\pi_{ij} = 1 - (1 - \psi)^{||N(i,j)||}$. Suppose $S_{0i} = 1$, and $S_{0j} = 0$. Then dyad (i, j) is observed, but $||N(i, j)||$ is unknown because it is unknown which k satisfy $y_{jk} = 1$. The link-tracing sampling structures for which nodal and dyadic sampling probabilities are observable are summarized in Table 1. For directed networks, we assume sampled nodes provide information on their out-arcs only, so that the design matrix is not symmetric and $D_{ij} = 1 \iff S_i = 1$.

Of the designs considered here, dyadic sampling probabilities are observable only for ego-centric samples, and never for link-tracing designs. Nodal sampling probabilities are also observable for ego-centric sampling, as well as for one-wave and total-wave link-tracing designs in undirected networks. Overall, this table presents strong limitations to the applicability of design-based methods requiring the knowledge of sampling probabilities to link-tracing designs. Note that this limitation is not specific to dyad-based network statistics. Estimation of triad-based network statistics such as a triad census would be subject to similar limitations. A Horvitz-Thompson style estimator would rely on a weighted sum

Table 1: Observable sampling probabilities under various sampling schemes for directed and undirected networks. Nodal and dyadic sampling probabilities are considered separately. “X” indicates observable sampling probabilities, while a blank indicates unobservable sampling probabilities.

Sampling	Nodal Probabilities π_i		Dyadic Probabilities π_{ij}	
Scheme	Undirected	Directed	Undirected	Directed
Ego-centric	X	X	X	X
One-Wave	X			
$k-$Wave, $1 < k < \infty$				
Total-Wave	X			

of observed triads, weighted according to sampling probabilities. Sampling probabilities for triads would be even more complex, as they would typically require sampling of two of the three nodes involved in an undirected case, and at least two of the three nodes in an directed case, depending on the triad census. Both of these sampling probabilities would not be possible to compute for link-tracing samples in which the degrees or in-degrees of some involved nodes are unobserved.

Not surprisingly, most of the work on design-based estimators for link-tracing samples has focused on the cases where sampling probabilities are observable: typically for one-wave or total-wave samples used to estimate population means of nodal covariates. Frank (2005) presents a good overview and extensive citations to this literature. See also Thompson and Collins (2002); Snijders (1992). Although examples tend to focus on instances where sampling probabilities are observable, the limited applicability of classical design-based methods in estimating structural network features based on link-tracing samples has not been emphasized in the literature.

In the absence of observable sampling probabilities, design-based inference requires a mechanism for estimating sampling probabilities. This is most often necessary in the context of out-of-design missing data, and addressed with approaches such as propensity scoring (Rosenbaum and Rubin, 1983), which rely on auxiliary information available for the full sampling frame to estimate unknown sampling probabilities. Link-tracing differs from the traditional context of such methods in that the sampling probabilities are unobserved even when the design is executed faithfully, and in that the unknown sampling probabilities result directly from the unobserved variable of interest. In particular, estimating unknown sampling probabilities is equivalent to estimating numbers of unobserved relations based on the observed relations. To do so, we must rely on a model relating the observed portions of the network structure to the unobserved portions. Lack of reliance on an assumed outcome model is a great advantage of design-based inference over model-based inference. By introducing a model to estimate sampling probabilities based on our outcome of interest, we re-introduce this reliance on model form, negating much of the advantage of design-based inference. Note however that the naive use of this approach has an ad-hoc flavor, while still requiring complex observation weights and variance estimators.

In the next section, we describe an alternative more flexible model-based approach to network inference based on link-tracing samples.

3.2 Model-based inference

Consider a parametric model for the random behavior of Y depending on a parameter $p-$vector η:

$$P_\eta(Y=y) \qquad \eta \in \Xi \tag{1}$$

In the model-based framework, if Y is completely observed inference for η can be based on the likelihood:

$$L[\eta|Y_{obs}] \propto P_\eta(Y=Y_{obs})$$

This situation has been considered in detail in Hunter and Handcock (2006) and the references therein. In the general case where Y may be only partially observed we can consider using the (so-called) *face-value likelihood* based solely on Y_{obs} :

$$L[\eta|Y_{obs}] \propto pr(Y_{obs}|\eta) = \int P_\eta(Y=y)dY_{mis}.$$

This ignores the additional information about η available in D. Inference for η and ψ should be based on all the available observed data, including the sampling design information. This likelihood is any function of η and ψ proportional to $pr(D, Y_{obs}|\eta,\psi)$:

$$L[\eta,\psi|Y_{obs},S] \propto pr(D,Y_{obs}|\eta,\psi) = \int pr(D|Y,\psi)P_\eta(Y=y)dY_{mis}$$

Thus the correct model is related to the complete data model through the sampling mechanism as well as the observed nodes and dyads.

In model-based inference, the sampling parameter ψ is a nuisance parameter, and modeling the complexity of the sampling mechanism along with the data structure adds a great deal of complexity. It is natural to ask when we might consider the simpler face-value likelihood:

$$L[\eta|Y_{obs}] \propto pr(Y_{obs}|\eta) = \int P_\eta(Y=y)dY_{mis},$$

which ignores the sampling mechanism.

In the context of missing data, Rubin (1976) introduced the concept of *ignorability* to specify when inference based on the face-value likelihood is efficient. We introduce the term *amenability* to represent the notion of ignorability for network sampling strategies within a model-based framework.

In many situations where models are used, the parameters $\eta \in \Xi$ and $\psi \in \Psi$ are *distinct*, in the sense that the joint parameter space of (η,ψ) is $\Psi \times \Xi$. If the design mechanism is adaptive and the parameters η and ψ are distinct:

$$\begin{aligned} & L[\eta,\psi|Y_{obs}=y_{obs},D=d] \\ \propto\ & pr(D=d|Y_{obs}=y_{obs},\psi)\int P_\eta(Y=y)dY_{mis} \\ \propto\ & L[\psi|D=d,Y_{obs}=y_{obs}]L[\eta|Y_{obs}=y_{obs}] \end{aligned}$$

Thus if the design mechanism is adaptive and the structural and sampling parameters are distinct, then the design mechanism is *ignorable* in the sense that the resulting likelihoods are proportional. When this condition is satisfied likelihood-based inference for η, as proposed here, is unaffected by the (possibly unknown) mechanism. This leads to the following definition and result.

Definition: Consider a design mechanism governed by parameter $\psi \in \Psi$ and a stochastic network model $P_\eta(Y = y)$ governed by parameter $\eta \in \Xi$. We call the design mechanism *amenable to the model* if the design mechanism is adaptive and the parameters ψ and η are distinct.

Result: Consider networks produced by the stochastic network model $P_\eta(Y = y)$ governed by parameter $\eta \in \Xi$ which are sampled by a design mechanism with parameter $\psi \in \Psi$ amenable to the model. Then the likelihood for η and ψ is

$$L[\eta, \psi | Y_{obs} = y_{obs}, D = d] \propto L[\psi | D = d, Y_{obs} = y_{obs}] L[\eta | Y_{obs} = y_{obs}]$$

Thus likelihood-based inference for η from $L[\eta, \psi | Y_{obs}, D]$ will be the same as likelihood-based inference for η based on $L[\eta | Y_{obs}]$.

This result shows for standard designs such as the ego-centric, single wave and multi-wave sampling designs in Section 2, likelihood-based inference can be based on the face-value likelihood $L[\eta | Y_{obs}]$. Explicitly, this is:

$$L[\eta | Y_{obs} = y_{obs}] \propto pr(Y_{obs} = y_{obs} | \eta) = \sum_{v\,:\, y_{obs} + v \in \mathcal{Y}} P_\eta(Y = y_{obs} + v)$$

where v has the same structure as Y_{mis} so that v_{ij} is undefined if $D_{ij} = 1$. Hence we can evaluate the likelihood by just enumerating the full data likelihood over all possible values for the missing data.

We may also wish to make inference about the design parameter ψ. The likelihood for ψ based on the observed data is any function of ψ proportional to $pr(D, Y_{obs} | \psi)$. For designs amenable to the model this is:

$$L[\psi | D, Y_{obs}] \propto pr(D | Y_{obs} = y_{obs}, \psi) \quad = \quad pr(D | Y_{obs} = y_{obs}, Y_{mis} = v, \psi)$$

for any compatible choice of y. Hence it can be computed directly.

4 Exponential Family Models for Networks

The models we consider for the random behavior of Y rely on a p-vector $g(Y)$ of statistics and a parameter vector $\eta \in R^p$. The canonical exponential family model is

$$P_\eta(Y = y) = \exp\{\eta^T g(y) - \kappa(\eta)\} \qquad y \in \mathcal{Y} \tag{2}$$

where $\exp\{\kappa(\eta)\} = \sum_{u \in \mathcal{Y}} \exp\{\eta^T g(u)\}$ is the familiar normalizing constant associated with an exponential family of distributions (Barndorff-Nielsen, 1978; Lehmann, 1983).

The range of network statistics that might be included in the $g(y)$ vector is vast — see Wasserman and Faust (1994) for the most comprehensive treatment of these statistics —

though we will consider only a few in this article. We allow the vector $g(y)$ to include covariate information about nodes or edges in the graph in addition to information derived directly from the matrix y itself.

There has been a lot of work on models of the form (2), to which we refer as exponential family random graph models or ERGMs for short.(We avoid the lengthier EFRGM, for "exponential family random graph models," both for the sake of brevity and because we consider some models in this article that should technically be called *curved* exponential families (Hunter and Handcock, 2006))

The normalizing constant is usually difficult to compute directly for Y containing large numbers of networks. Inference for this class of models was considered in the seminal paper by Geyer and Thompson (1992), building on the methods of Frank and Strauss (1986) and the above cited papers. Until recently, inference for social networks models has relied on maximum pseudolikelihood estimation (Besag, 1974; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson, 1992). Geyer and Thompson (1992) proposed a stochastic algorithm to approximate maximum likelihood estimates for model (2) among other models; this Markov chain Monte Carlo (MCMC) approach forms the basis of the method described in this article. The development of these methods for social network data has been considered by Corander et al. (1998); Crouch et al. (1998); Snijders (2002); Handcock (2002); Corander et al. (2002); Hunter and Handcock (2006).

4.1 Model-based inference for ERGM

In this section we consider likelihood inference for η in the case where $Y = Y_{obs} + Y_{mis}$ is possibly only partially observed.

As this may entail a large number of terms, we can approximate the likelihood by using the MCMC trick of randomly sampling from the space of possible values of the missing data and taking the mean. Alternatively consider the conditional distribution of Y given Y_{obs} :

$$P_\eta(Y_{mis} = y | Y_{obs} = y_{obs}) \ = \ \exp\left[\eta^T g(y + y_{obs}) - \kappa(\eta | y_{obs})\right] \qquad\qquad y \in \mathcal{Y}$$

where $\exp\left[\kappa(\eta|y_{obs})\right] = \sum\limits_{u: u+y_{obs} \in \mathrm{Y}} \exp\left[\eta^T g(u + y_{obs})\right]$. This formula gives a simple way to sample from the conditional distribution and hence produce multiple imputations of the full data.

Also note that

$$L[\eta | Y_{obs} = y_{obs}] \propto \exp\left[\kappa(\eta|y_{obs}) - \kappa(\eta)\right]$$

which can then be estimated by MCMC samples: the first term by a chain on the complete data and the second by a chain conditional on y_{obs} . So the sampled data situation is only a little bit harder than the complete data case.

5 Two-wave link-tracing samples from a Legal Network

In this section we investigate the effect of network sampling on estimation by comparing network samples to the situation where we observe the complete network. The Lazega

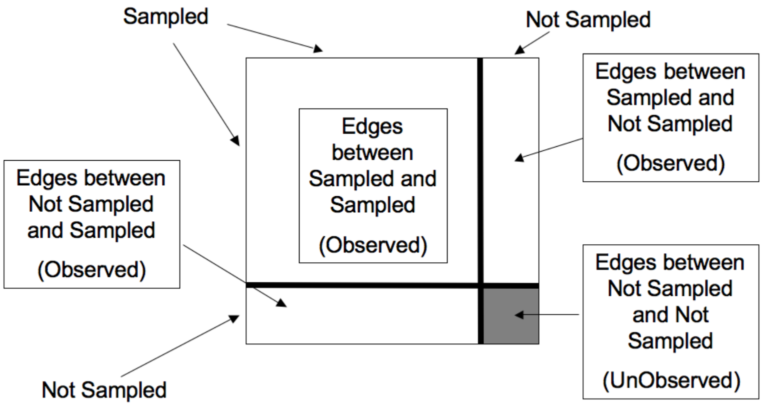
Sampled
Not Sampled
Edges between Sampled and Not Sampled (Observed)
Edges between Sampled and Sampled (Observed)
Edges between Not Sampled and Sampled (Observed)
Edges between Not Sampled and Not Sampled (UnObserved)
Not Sampled

Figure 1: Schematic depiction of sampled and unobserved arc data when the sampling is over an undirected network.

(2001) undirected collaboration network of 36 law firm partners is used as the basis for the study. In assessing the effect of sampling on model fit we start with a well fitting model for the data. We consider Model 2 in Hunter and Handcock (2006). The structural parameters, related to network statistics, are the number of edges (essentially the density) and the geometrically weighted edgewise shared partner statistic (denoted by GWESP), a measure of the transitivity structure in the network. Two nodal attributes are used: seniority (ranknumber/36) and practice (corporate or litigation). Three dyadic homophily attributes are used: practice, gender (3 of the 36 lawyers are female) and office (3 different locations of different size). The model has been slightly reparameterized from Hunter and Handcock (2006) by replacing the alternating k-triangle term with the GWESP term. The scale parameter for the GWESP term is fixed at its optimal value (0.7781). (See Hunter and Handcock, 2006, for details). A summary of the MLE parameters used is given in column two of Table 2. Note that we are taking these parameters as "truth" and considering data produced by sampling from this network.

We construct all possible datasets produced by a two-wave link-tracing starting from two randomly chosen nodes (the "seeds"). This adaptive design is amenable to the model. As there are 36 partners and the sample is deterministic given the seeds, there are $\binom{36}{2} = 630$ possible data sets. The number of actors in each dataset varies from just 2 to all 36 depending on the degree of connectedness of the seeds. The data pattern is shown in Figure 5. Consider a partition of the sampled from the non-sampled and the corresponding 2×2 blocking of the sociomatrix, with the four blocks representing dyads from sampled and non-sampled to sampled and non-sampled. The complete data consists of the full sociomatrix. The first three blocks contain the observed data, the dyads involving at least one respondent, and the last block contains the unobserved data, those between the non-sampled.

For each of these samples we use the methods of Section 4.1 to estimate the parameters. We can then compare them to the MLE for the complete dataset. For these networks, the MLEs are obtained using `statnet` (Handcock et al., 2003), both for the natural parametrization and for the mean value parameterization (see Handcock, 2003).

The mean value parameters are a function of the natural parameters, specifically the expected values of the sufficient statistics given the values of the natural parameters.

There are two isolates, that is nodes with no relations. If these two are selected as the two seeds, only 69 of the 630 dyads are observed. There are also two pairs of seeds where only 5 partners appear in at least one of the waves (corresponding to observing 165 (26%) dyads). Due to the smaller sample sizes, the estimates from these 3 samples are quite variable compared to the other 627. Note that the issue here is the number of dyads sampled and their relationship rather than the percentage sampled. The sampler will not know that these samples are extreme and so an evaluation of the sampling process should include them. However the sampler may be concerned about the (known) small sample size. It is unlikely any researcher drawing a link-tracing sample including only two isolated nodes will proceed with analysis of that sample. It is more likely the sampler would proceed with a sample including only five partners, but still doubtful. In any case we include population-level comparisons both including these extremely small samples and excluding them.

One way to assess the effect of the link-tracing design is to compare the estimates from the sampled data to that of the complete data. As a measure of the difference between the estimates in the metric of the model, we use the Kullback-Leibler divergence from the model implied by the complete data estimate to that of the sampled data estimate. Recall that the Kullback-Leibler divergence of a distribution with probability mass function p from the distribution with probability mass function q is

$$E_q[\log(q) - \log(p)]$$

Let η and ξ be alternative parameters for the model (2). The Kullback-Leibler divergence, $KL(\xi, \eta)$, of the ERGM with parameter η from the ERGM with parameter ξ is:

$$\begin{aligned} E_\xi\left[\log\left(\frac{P_\xi(Y=y)}{P_\eta(Y=y)}\right)\right] &= \sum_{y\in\mathcal{Y}} \log\left(\frac{P_\xi(Y=y)}{P_\eta(Y=y)}\right) P_\xi(Y=y) \\ &= \sum_{y\in\mathcal{Y}} (\xi-\eta)^T y P_\xi(Y=y) + \kappa(\eta) - \kappa(\xi) \\ &= (\xi-\eta)^T E_\xi[g(Y)] + \kappa(\eta) - \kappa(\xi) \end{aligned}$$

If ξ is the complete data MLE then $E_\xi[g(Y)] = g(Y_{obs})$ are the observed statistics (given in column 2 of Table 3). The divergence can be easily computed using the MCMC algorithms of Section 4.1.

Figure 5 plots the Kullback-Leibler divergence of the MLEs based on the 627 samples from the complete data MLE. The Kullback-Leibler divergence of the three extreme samples are 14 to 18 have not been plotted to reduce the vertical scale. The horizontal axis is the number of observed dyads in the sample. The plot indicates how the information in the data about the complete data MLE approaches that of the complete data as the number of sampled dyads approaches the full number. The key feature of this figure is the *variation* in information content among samples of the same size especially for the smaller sample sizes. Different seeds lead to samples that tell us different things about the model even when the numbers of partners surveyed is the same.

For more specific information on the individual estimates, we can compute the bias of the estimates based on the samples as the mean difference between the parameter estimates from

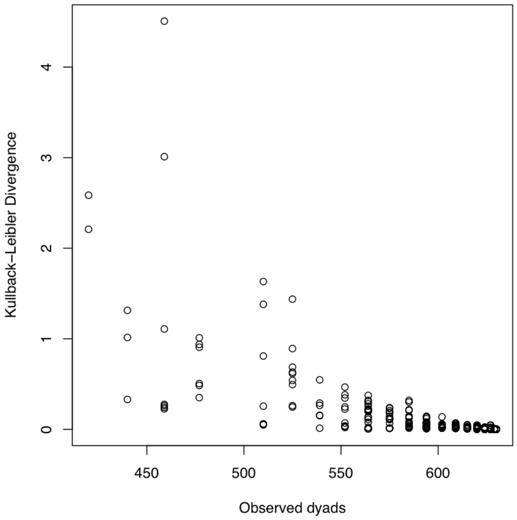

Figure 2: Kullback-Leibler divergence of the MLEs based on the samples compared to the complete data MLE. As the number of dyads sampled increases the information content of the samples approaches that of the complete data. The information loss for the majority of samples is modest.

the samples and that of the complete network. The root mean squared error (RMSE) is the square-root of the mean of the squared difference between the parameter estimates from each sample and the complete data estimates. The efficiency loss of the sampled estimate is the ratio of the mean squared error and the variance of the sampling distribution of the estimate based on the full data. This standardizes the error in the sampled estimates by the variation in the complete data estimates remaining in the complete data. We also complete a similar comparison of the estimates under the alternative mean value parametrization (Handcock 2003).

The properties of the original model's natural parameter estimates are summarized in Table 2. The bias and root mean squared error are presented in percentages of the complete data parameter estimates.

When the three extreme samples are excluded, the bias is very small and the RMSE is modest. The efficiency loss is 1-2% on average. Note that these population-average figures obscure the variation in loss over individual samples apparent in Figure 5. A consideration of all samples, leaves the bias small but leads to an increase in the RMSE. The efficiency losses are also substantially increased, especially those of the edges and GWESP terms. For these the errors are respectively 21.9% and 11.1% of the variation in the complete data. However, as we have seem, much of this is from the few extremely bad samples.

Table 3 is the mean value parameterization analog of Table 2. As these are on the same measurement scale as the statistics they are easier to interpret. Again we see that the estimates are approximately unbiased whether the extreme values are included or excluded. However, for these parameters the efficiency loss is small for overall samples.

Table 2: Bias and Root Mean Squared Error (RMSE) of natural parameter MLE based on two-wave samples as percentages of true parameter values and efficiency losses

	complete	Excluding worst 3 of 630			All 630 possible samples		
natural	data	bias	RMSE	efficiency	bias	RMSE	efficiency
parameter	value	(%)	(%)	(%)	(%)	(%)	loss (%)
structural							
edges	-6.51	0.2	1.0	1.3	0.3	4.1	21.9
GWESP	0.90	0.9	2.5	2.4	0.9	5.4	11.1
nodal							
seniority	0.85	0.3	3.1	1.2	0.5	5.1	3.4
practice	0.41	0.2	3.7	1.7	0.2	6.6	5.4
homophily							
practice	0.76	0.7	3.9	2.3	0.9	5.9	5.3
gender	0.70	1.0	4.4	1.5	0.8	6.2	2.9
office	1.15	0.8	2.7	2.5	0.6	4.9	8.3

Table 3: Bias and Root Mean Squared Error (RMSE) of mean value parameter MLE based on two-wave samples as percentages of true parameter values and efficiencies

	complete	Excluding worst 3 of 630			All 630 possible samples		
natural	data	bias	RMSE	efficiency	bias	RMSE	efficiency
parameter	value	(%)	(%)	(%)	(%)	(%)	loss (%)
structural							
edges	115.00	0.4	2.0	1.8	0.4	2.0	1.8
GWESP	190.31	0.3	2.6	1.6	0.4	2.8	1.9
nodal							
seniority	130.19	0.0	0.1	1.4	0.0	0.1	1.4
practice	129.00	0.1	2.0	1.7	0.2	2.6	3.4
homophily							
practice	72.00	0.1	1.8	1.7	0.1	2.0	1.7
gender	99.00	0.5	2.1	1.8	0.5	2.1	1.8
office	85.00	0.7	2.6	3.0	0.7	2.7	3.0

6 Discussion

In this paper we give a concise and systematic statistical framework for dealing with partially observed network data resulting from a designed sample. The framework includes, but is not restricted to, adaptive network sampling designs. We present a definition of a network design which is amenable to a given model and a result on likelihood-based inference under such designs.

An important simple result of this framework is that sampled networks are not "biased" but can be representative if analyzed correctly. Many authors have confused the ideas of simple random sampling of the dyads with representative designs. The results of this paper indicate that simple random sampling is not necessary for valid inference. In fact, the most commonly used designs can be easily taken into account. Hence, despite their form, inference for adaptive network sampled information is tractable.

We have also shown that model-based inference from an adaptive network sample can be conducted using a complete network model. We have shown that such inference is both principled and practical. The likelihood framework naturally accommodates standard sampling mechanisms. Note that in a design-based frame, principled inference would require a great deal of effort to precisely characterize the sampling mechanism. The result that link-tracing designs are adaptive and can be analyzed with likelihood based methods is very valuable in practice as these designs have previously not been analyzed with general ERG (or similar) models.

In our application we show that an adaptive network sampling of a collaboration network can lead to effective estimates of the model parameters in the vast majority of cases. We find that the MLEs from the samples have only modest bias (compared to the complete data estimate) and an error that only increases slowly with the number of unobserved dyads. We also show that the information content of the sample (with respect to the model), varies greatly even for samples of the same size. For conventional samples of i.i.d. random variables, the Fisher information is simply proportional to the sample size. In the network setting with dependence-terms, however, the Fisher information will depend on the specific set of nodes and dyads sampled. For example, the information component corresponding to the GWESP term in the example will be larger for samples in which more pairs of nodes joined by edges are sampled, as GWESP applies only to pairs of nodes joined by edges. If no such dyads were sampled, there would be no information in the sample about the propensity for dyads sharing edges to have relations in common.

In practice the sample is a result of a combination of the design mechanism and *out-of-design* mechanism. The design mechanism is that part of the observation process under the control of the surveyor. When adaptive designs are executed faithfully, the unknown dyads are assumed to be intentionally unobserved, or missing by design. The definition of control may be extended by allowing the design to depend on unknown factors, such as the unrecorded values of variables used for stratification. The out-of-design mechanism is the non-intentional non-observation of network information (e.g., due to the failure to report links, incomplete measurement of links and attrition from longitudinal surveys). This is also referred to, in general, as the *non-response mechanism.* We consider the joint effect of sampling and missing data in a companion paper (Handcock and Gile, 2007).

We will make available the code used in this study on the `statnet` website (Handcock et al., 2003). `statnet` is an open-source software suite for network modeling and is written for the `R` environment.

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* New York: Wiley.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Corander, J., K. Dahmström, and P. Dahmström (1998). Maximum likelihood estimation for markov graphs. Research report, Department of Statistics, University of Stockholm.

Corander, J., K. Dahmstrom, and P. Dahmstrom (2002). Maximum likelihood estimation for exponential random graph models. In J. Hagberg (Ed.), *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank*, pp. 1–17. Stockholm: University of Stockholm, Department of Statistics.

Crouch, B., S. Wasserman, and F. Trachtenberg (1998). Markov chain monte carlo maximum likelihood estimation for p^* social network models. In *Paper presented at the XVIII International Sunbelt Social Network Conference in Sitges, Spain.*

Frank, O. (2005). Network sampling and model fitting. In J. S. P. Carrington and S. S. Wasserman (Eds.), *Models and Methods in Social Network Analysis*, pp. in press. Cambridge: Cambridge University Press.

Frank, O. and D. Strauss (1986). Markov graphs. *Journal of the American Statistical Association 81*(395), 832–842.

Geyer, C. J. and E. A. Thompson (1992). Constrained monte carlo maximum likelihood calculations (with discussion). *Journal of the Royal Statistical Society, Series B 54*, 657–699.

Handcock, M. S. (2002). Degeneracy and inference for social network models. In *Paper presented at the Sunbelt XXII International Social Network Conference in New Orleans, LA.*

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Working paper #39, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M. S. and K. Gile (2007). Modeling social networks with sampled or missing data. Working paper, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2003). *statnet: An R package for the Statistical Modeling of Social Networks.* http://statnetproject.org.

Hunter, D. R. and M. S. Handcock (2006, September). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics 15*(3), 565–583.

Lazega, E. (2001). *The collegial phenomenon: the social mechanisms of cooperation among peers in a corporate law partnership.* Oxford: Oxford University Press.

Lehmann, E. L. (1983). *Theory of Point Estimation.* New York, NY: John Wiley.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Snijders, T. A. B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin Methodologie Sociologique 36*, 59–70.

Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure 3*(2).

Strauss, D. and M. Ikeda (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association 85*, 204–212.

Thompson, S. K. and L. M. Collins (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence 68*, S57–S67.

Thompson, S. K. and O. Frank (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology 26*, 87–98.

Thompson, S. K. and G. A. F. Seber (1996). *Adaptive sampling*. New York: Wiley.

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