Modeling Social Networks with Sampled or Missing Data [1]

Working Paper no. 75
Center for Statistics and the Social Sciences
University of Washington

Mark S. Handcock
Krista Gile
University of Washington, Seattle

June 11, 2002; Revised April 28, 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 implications of these relations. In studies of social networks recent emphasis has been placed on random graph models where the nodes usually represent individual social actors and the edges represent a specified relationship between the actors.

Most inference for models for social networks assumes that the presence or absence of all links in the network are completely observed, that the information is completely reliable and 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 and consider the various forms of out-of-design missing data. 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 motivated and illustrate the ideas by analyzing the effect of link-tracing sampling designs on a collaboration network and by an analysis of social relations from the National Longitudinal Study of Adolescent Health subject to missing data.

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 regular pattern of the relations between nodes. Networks can be used to describe such diverse ideas such as the behavior of epidemics, the interconnectedness of the corporate boards, and network of genetic regulatory interactions. In social network applications, the nodes in a graph 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 this induces.

Social network data typically consist of a set of n actors and a relational tie $Y_{i,j}$, measured on each possible *dyad* $\{i, j\}$, an ordered pair of actors $i, j = 1, \ldots, n$. In the most simple cases, $Y_{i,j}$ 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 is often represented by an $n \times n$ sociomatrix Y. 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 $\{(i, j) : Y_{i,j} = 1\}$. When (i, j) is in the edge set we write $i \to j$. For many networks the relations are undirected in the sense that $\{Y_{i,j} = Y_{j,i}, i, j = 1, \ldots, n\}$. To simplify the presentation, we focus on directed binary relations.

The structure of the relations is usually dependent on the attributes of the actors. For example, for most social relations the likelihood of a relationship is a function of the age, gender, geography and race of the individuals. Homophily of attributes is often associated with increased propensity of a relationship (McPherson, Smith-Lovin, and Cook, 2001), although the effect may be reversed (e.g., gender and sexual relationships). In addition to exogenous attributes of the actors, relationships are influenced by endogenous attributes such as their positions in the network (White, Boorman, and Breiger 1976). 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. An important aspect of these methods is that they adaptively exploit the components of the networks as they are observed to guide the sampling. These adaptive designs allow for more efficient sampling than conventional sampling designs. We consider such design in Section 3.

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

The statistical modeling of networks has a long history. Holland and Leinhardt (1981) appear to be the first to propose log-linear models for social networks. Their models resulted in each dyad — by which we mean each pair of actors — having edges independently of every other dyad. Frank and Strauss (1986) generalized to the case in which dyads exhibit a form of Markovian dependence: Two dyads are dependent, conditional on the rest of the graph, only when they share a node. Such exponentially parametrized random graph models have connections to a broad array of literatures in many fields, such as spatial statistics, statistical exponential families, and statistical physics (Geyer and Thompson, 1992).

In this paper we develop a theoretical framework of missing information and sampling of networks. These extend those in the fundamental work of Thompson and Frank (2000).

In Section 2 we present ...

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 (1995) and Thompson and Frank (2000). Suppose there are q (exogenous) covariates on the ordered pair of actors which we denote by the $n \times n \times q$ array $X = \{X_{i,j}\}$ $i, j = 1, \ldots, n$ where $X_{i,j}$ is the $q-$vector covariates associated with the $\{i, j\}$, dyad. Let $\mathcal{X}$ denote the sample space of covariates and $\mathcal{Y}(x)$ the set of possible networks on the n actors with a given set of covariates $x \in \mathcal{X}$. Let $(\mathcal{Y}, \mathcal{X})$ denote the joint set of networks and covariates. For a given network $y \in \mathcal{Y}(\mathcal{X})$ with covariates $x \in \mathcal{X}$, denote a sample from the network and covariates as $s \subset \{y, x\}$. The sample is selected by 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 (e.g., a survey using conventional, egocentric, snowball or other link-tracing sampling). 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. However in the usual case the stochastic mechanism of the design is known completely, or up to a parameter ψ. Let $p(s|y, x; \psi)$ denote the probability of the design mechanism selecting sample s given a network y and covariates x. 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.*

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 $p(s|y,x;\psi) = p(\psi)\ \forall x, y \in (\mathcal{Y}, \mathcal{X})$. A simple example of a conventional design mechanisms for networks is simple random sampling of a subset of the actors, followed by complete observation of the pairs originating from those actors. A complete census of the network and covariates is another. More complex examples are designs using probability sampling of pairs and auxiliary variables.

We call a design mechanism *adaptive* if it uses information collected during the survey to direct subsequent sampling of actors, but the mechanism depends only on the observed information. Specifically, a design is adaptive if: $p(s|y,x;\psi) = p(s;\psi)\ \forall x, y \in (\mathcal{Y}, \mathcal{X})$. The definitions of conventional and adaptive designs can be refined to allow this condition to hold for a subset of $(\mathcal{Y}, \mathcal{X})$.

3 Model-based Sampling

In the "design-based" framework $\{y, x\}$ represents the population and interest focuses on characterizing x based on partial observation. Under the "model-based" framework X is stochastic and is a realization from a stochastic process depending on a parameter $\boldsymbol{\eta}$. Here interest focuses on $\boldsymbol{\eta}$ which characterizes the mechanism that produced the complete graph X. The model may also be used to guide design-based inference (Sarndal, Swensson, and Wretman 1992).

For notational simplicity, we focus on the sampling of the network and suppress reference to the covariates. We return to these in Section 4.2.

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} pair was intentionally not sampled and 1 if the element was chosen to be sampled. We shall refer to D as the *network design mechanism.* We shall refer to realizations of D as the *design matrix* and the distribution of D as the *design mechanism.* The design mechanism is usually related to the structure of the graph so we posit a model for it,

$$pr(D|Y),$$

which depends on Y and typically the exogenous attributes X which we suppress.

Denote the observed part of the complete graph Y by $Y_{obs} = \{Y_{ij} : D_{ij} = 1\}$ and the unobserved part by $Y_{mis} = \{Y_{ij} : D_{ij} = 0\}$. What we observe is then the *observed data:* $\{Y_{obs}, D\}$, in contrast to the *complete data:* $\{Y_{obs}, Y_{mis}, D\}$. We will write the complete graph $Y = \{Y_{obs}, Y_{mis}\}$.

Formally, we can also represent Y_{obs} as $n \times n$ matrix indicating the corresponding element of Y if it is observed, and undefined if it are not:

$$Y_{obs,ij} = \begin{cases} Y_{ij} & \text{if } D_{ij} = 1 \\ ? & \text{if } D_{ij} = 0 \end{cases} .$$

The reverse is true of Y_{mis} :

$$Y_{mis,ij} = \begin{cases} Y_{ij} & \text{if } D_{ij} = 0 \\ ? & \text{if } D_{ij} = 1 \end{cases} .$$

In addition, if we make the convention that a number plus or multiplied by an undefined (i.e. "?") is the number, we have $Y = Y_{obs} + Y_{mis}$.

3.1 Example: Ego-Centric designs

For example, 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 order 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 . Let 1 be the binary $n-$vector of 1s, and denote by $[y]$ the vector-valued function that is 1 if the corresponding element of y is logically true, and 0 otherwise.

Let s_0 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 observation matrix (i.e. $s_0 = [D1 = n1]$) . Then $pr(s_o = s|Y, \psi) = \psi^{1^T s}(1 - \psi)^{n-1^T s} \ \ s \in \{0, 1\}^n$. If the ith element of s_0 is 1 then all elements in the ith row and column of D are 1 (and 0 otherwise). 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.

3.2 Example: One-wave link tracing design

Let s_0 denote the indicator 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 observation matrix is given by $1s^T - s1^T - ss^T \quad s \in \{0,1\}^n$. Note that $s_1 = [Ys_0 > 0]$ is derivable from s and Y. Hence

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

for

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

3.3 Example: Multi-wave link tracing design

Consider a *multi-wave link tracing design* or *complete wave snowball design* in which the complete set of partners of the kth wave are enrolled.

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, ..., s_k the indicator for the added individuals in wave k not in the prior samples. Then the whole sample of individuals is $s = s_0 + s_1 + ... + s_k$. As in the ego-centric design the observation matrix is given by $1s^T - s1^T - ss^T \quad s \in \{0,1\}^n$. Note that $s_m = [Ys_{m-1} > 0]\,, \quad m = 1, ..., k$ is derivable from s_0 and Y .

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

for

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

Note that $s_m = [Ys_{m-1} > 0] = [Y_{obs}s_{m-1} > 0]\,, \quad m = 1, ..., k$ so that the individuals selected in the successive waves of a complete wave snowball sample depend only on the observed part of the graph.

3.4 Design-based inference for Y_{mis}

In the design-based framework, we wish to make inference about ψ and also about the unknown values Y_{mis}. The likelihood for ψ based on the observed data is any function of ψ proportional to $pr(D, Y_{obs}|\psi)$:

$$L[\psi|D, Y_{obs}] \propto pr(D, Y_{obs}|\psi) \quad = \quad \int pr(D|Y, \psi) dY_{mis}.$$

4 Model-based inference for $\boldsymbol{\eta}$

Consider a parametric model for the random behavior of Y depending on a parameter $p-$vector $\boldsymbol{\eta}$:

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

In the model-based framework, if Y is completely observed inference for $\boldsymbol{\eta}$ can be based on the likelihood:

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

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} :

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

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

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

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

4.1 Ignorability of the Sampling Mechanism

It is natural to ask when inference for $\boldsymbol{\eta}$ should be based on the observed data likelihood and when it can be based on the simpler face-value likelihood which ignores the sampling mechanism. For many surveys the sampling mechanism satisfies

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

a condition called "*missing at random*" by Rubin (1976). 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 graph, but that this relationship can be explained by the data that is observed. The observed part of the data are usually vital to this equality.

In many situations where models are used, the parameters $\boldsymbol{\eta}$ and ψ are *distinct*, in the sense that the joint parameter space of $(\boldsymbol{\eta}, \psi)$ is the product of the parameter space of $\boldsymbol{\eta}$

and the parameter space of ψ. If the sampling mechanism is missing at random and the parameters $\boldsymbol{\eta}$ and ψ are distinct:

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

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

Thus if the sampling mechanism is missing at random then the sampling mechanism is *ignorable* in the sense that the resulting likelihoods are proportional.

The same holds true for the design-based inference: the model for the graph (but not the graph itself!) is ignorable for inferences about the sampling mechanism parameter ψ.

When this condition is satisfied likelihood-based inference for $\boldsymbol{\eta}$, as proposed here, is unaffected by the (possibly unknown) sampling mechanism.

4.2 Inference for Adaptive Network Sampling Designs

In this section we consider likelihood inference for conventional and adaptive network sampling designs. For completeness we do not suppress the reference to the covariates X. Denote the observed part of the covariate array X by $X_{obs} = \{X_{ij} : D_{ij} = 1\}$ and the unobserved part by $X_{mis} = \{X_{ij} : D_{ij} = 0\}$. The full observed data is then $\{Y_{obs}, X_{obs}, D\}$. Suppose the sampling design is conventional. Then

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

and the design is ignorable. More generally, if the sampling design is adaptive:

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

so these designs are ignorable. Specifically:

Result: Suppose that the network data (Y, X) follows a stochastic process $P_{\boldsymbol{\eta}}(Y = y, X = x)$ governed by a $q-$vector parameter $\boldsymbol{\eta} \in \Xi$. Suppose the network sampling design mechanism, D, is governed by a parameter ψ and produces the data (y_{obs}, x_{obs}). If the network sampling design mechanism is adaptive and the parameters $\boldsymbol{\eta}$ and ψ are distinct then the network sampling design mechanism is ignorable. Consequently the likelihood for $\boldsymbol{\eta}$ and ψ is

$$\begin{aligned} & L[\boldsymbol{\eta}, \psi | Y_{obs} = y_{obs}, X_{obs} = x_{obs}, D = d] \\ \propto \quad & L[\psi | D = d, Y_{obs} = y_{obs}, X_{obs} = x_{obs}] L[\boldsymbol{\eta} | Y_{obs} = y_{obs}, X_{obs} = x_{obs}] \end{aligned}$$

Henceforth we assume the parameters $\boldsymbol{\eta}$ and ψ are distinct. The result shows that the ego-centric, single wave and multi-wave sampling are ignorable, and likelihood-based inference can be based on the face-value likelihood $L[\boldsymbol{\eta}|Y_{obs}]$. Explicitly, this is:

$$\begin{aligned} & L[\boldsymbol{\eta}|Y_{obs} = y_{obs}, X_{obs} = x_{obs}] \\ \propto \quad & pr(Y_{obs} = y_{obs}, X_{obs} = x_{obs}|\boldsymbol{\eta}) \\ = \quad & \sum_{(x,y)\,:\,(y_{obs}+y,\ x_{obs}+x)\,\in\,(\mathcal{X},\mathcal{Y})} P_{\boldsymbol{\eta}}(Y = y_{obs} + y, X = x_{obs} + x) \end{aligned}$$

where y and x have the same structure as Y_{mis} and X_{mis} while y_{ij} and x_{ij} are 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.

5 Out-of-design mechanism

The *out-of-design mechanism* or *non-response mechanism* refers to non-intentional non-observation of dyads. Some sources of sampling factors that lead to out-of-design missing data are:

- Non-random sample of respondents – the initial sampling design was non-random or in some way a convenience sample
- Inaccuracy of reported links – contact information does not allow identification or contacts are miss-identified
- Dead-ends due to mobility of contacts: These occur when contacts are untraceable even when accurately reported by the respondent.
- Failure to complete full link-tracing in the design
- Failure to report all links – some contacts are not disclosed
- Missing responses on covariates
- Inaccuracy of reported covariates

Such procedures can be considered as "sampling" of the data but with a mechanism that is not completely under the control of the surveyor. However we can consider some forms of missing data in the same manner as we consider adaptive network designs. Specifically, we cause the traditional notion of ignorability (Rubin, 1976). Suppose the out-of-design mechanism, O, is governed by a parameter ψ and produces the data $s = (y_{obs}, x_{obs})$. We call

O ignorable if $p(s|y,x;\psi) = p(s;\psi)\ \forall x, y \in (\mathcal{Y}, \mathcal{X})$. as the mechanism does not depend on the unobserved information given the observed information. We then can restate the following classic result for this circumstance:

Result: Suppose that the network data (Y, X) follows a stochastic process $P_{\boldsymbol{\eta}}(Y = y, X = x)$ governed by a $q-$vector parameter $\boldsymbol{\eta} \in \Xi$. If the out-of-design design mechanism is ignorable and the parameters $\boldsymbol{\eta}$ and ψ are distinct then the likelihood for $\boldsymbol{\eta}$ and ψ is

$$\begin{aligned} & L[\boldsymbol{\eta}, \psi | Y_{obs} = y_{obs}, X_{obs} = x_{obs}, D = d] \\ \propto \quad & L[\psi | D = d, Y_{obs} = y_{obs}, X_{obs} = x_{obs}] L[\boldsymbol{\eta} | Y_{obs} = y_{obs}, X_{obs} = x_{obs}] \end{aligned}$$

In the next section we consider a rich class of parametric models for Y that can be used under both adaptive sampled and ignorable network designs.

6 Exponential Family Models for Networks

The models we consider for the random behavior of Y rely on a p-vector $\mathbf{Z}(Y|X = x)$ of statistics and a parameter vector $\boldsymbol{\eta} \in R^p$. The canonical exponential family model is

$$P_{\boldsymbol{\eta}}(Y = y | X = x) = \exp\{\boldsymbol{\eta} \cdot \mathbf{Z}(y; x) - \psi(\boldsymbol{\eta}; x)\}, \tag{2}$$

where

$$\exp\{\psi(\boldsymbol{\eta}; x)\} = \sum_u \exp\{\boldsymbol{\eta} \cdot \mathbf{Z}(u; x)\} \tag{3}$$

is the familiar normalizing constant associated with an exponential family of distributions (Barndorff-Nielsen 1978; Lehmann, 1983). The sum in (3) is taken over the whole sample space, which presents a very important problem in most applications: A sample space consisting of all possible directed graphs on n nodes contains $\exp\{n(n-1)\log 2\}$ elements, an astronomically large number even for moderately sized n of, say, 20. Thus, for most applications it is impossible even to evaluate the likelihood function for a particular $\boldsymbol{\eta}$.

The range of network statistics that might be included in the $\mathbf{Z}(y; x)$ 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 $\mathbf{Z}(y; x)$ to include covariate information about nodes or edges in the graph in addition to information derived directly from the matrix y itself. Thus, $\mathbf{Z}(y; x)$ should be viewed as a function not only of y, but also potentially of the exogenous covariates x. For example, if each node is a person, $\mathbf{Z}(y; x)$ might include the total number of edges between individuals of the same gender, which is a function of both the graph y and the exogenous nodal covariate gender. For

notational simplicity, we prefer to allow the dependence of $\mathbf{Z}$ on exogenous covariates to be implicit rather than explicitly indicated by the notation.

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.) Holland and Leinhardt (1981) appear to be the first to propose a specific case of model (2) in the literature. Their model, which they called the p_1 model, resulted in each dyad — by which we mean each pair of nodes — having edges independently of every other dyad. Based on developments in spatial statistics (Besag 1974), Frank and Strauss (1986) generalized to the case in which dyads exhibit a kind of Markovian dependence: Two dyads are dependent, conditional on the rest of the graph, only when they share a node. Frank (1991) mentioned the application of model (2) to social networks in its full generality. This was pursued in depth by Wasserman and Pattison (1996). In honor of Holland and Leinhardt's p_1 model, they referred to model (2) as $p*$ ($p-$star), a name that has been widely applied to ERGMs in the social networks literature.

We note that the model (2) can be thought of as simply a parametrization of the set of possible model for the network, as they can represent any finite random graph model by appropriate choice of $\mathbf{Z}$.

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).

6.1 Model-based inference for ERGM

In this section we consider likelihood inference for $\boldsymbol{\eta}$ 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} :

$$P_{\boldsymbol{\eta}}(Y_{mis} = y | Y_{obs} = y_{obs}) \;=\; \frac{\exp\left[\boldsymbol{\eta}\cdot\mathbf{Z}(y + y_{obs}; x)\right]}{\mathrm{c}(\boldsymbol{\eta}|y_{obs})} \qquad y \in \mathcal{Y}$$

where $\mathrm{c}(\boldsymbol{\eta}|y_{obs}) = \sum_{y' : y'+y_{obs} \in \mathrm{Y}} \exp\left[\boldsymbol{\eta}\cdot\mathbf{Z}(y' + y_{obs}; y)\right]$. This formula gives a simple way to sample from the conditional distribution and hence produce multiple imputations of the full data.

Also note that

$$L[\boldsymbol{\eta}|Y_{obs} = y_{obs}] \propto \frac{\mathrm{c}(\boldsymbol{\eta}|y_{obs})}{\mathrm{c}(\boldsymbol{\eta})}$$

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

Thompson and Frank (2000) show that most link-tracing designs are missing at random, in this sense. In general, some aspects of the sampling mechanism may need to be modeled – for a review, see Schafer (1997).

In practice, however, any local network design will likely be subject to out-of-design missing and non-reliable data. For example, a common problem in network studies is that respondents over or under-report their partnerships. As a simple model for this, suppose that individual links (non-links) are independently erroneously reported as non-links (links) with probability $\alpha_1(\alpha_0)$. Hence we have non-reliable values of the sampled dyads where the reliability depends on the value of the dyad. The out-of-design mechanism is:

$$pr(R = r|Y, \alpha_0, \alpha_1) = \alpha_1^{[Yd \neq r]}(1-\alpha_1)^{[Yd=r]}\alpha_0^{[(11^T - Y)d \neq r]}(1-\alpha_0)^{[(11^T - Y)d=r]}$$

for $d = 1s^T - s1^T - ss^T \;\; s \in \{0,1\}^n$. The sampling mechanism is then

$$pr(D = d|Y, \psi, \alpha_0, \alpha_1) = \sum_{\substack{d\,:\,d \geq o \\ d = 1s^T - s1^T - ss^T}} pr(D = d|Y, \psi) pr(R = o|Y, D = d, \alpha_0, \alpha_1)$$

for $s \in \{0,1\}^n$. If the two error probabilities are the same (i.e. $\alpha_0 = \alpha_1$) the sampling mechanism, like the local network design, is ignorable. In general, however, this sampling mechanism is non-ignorable.

This is a simple but practical example of a model for a sampling mechanism which is non-ignorable due to out-of-design missing data, even though the underlying design is ignorable.

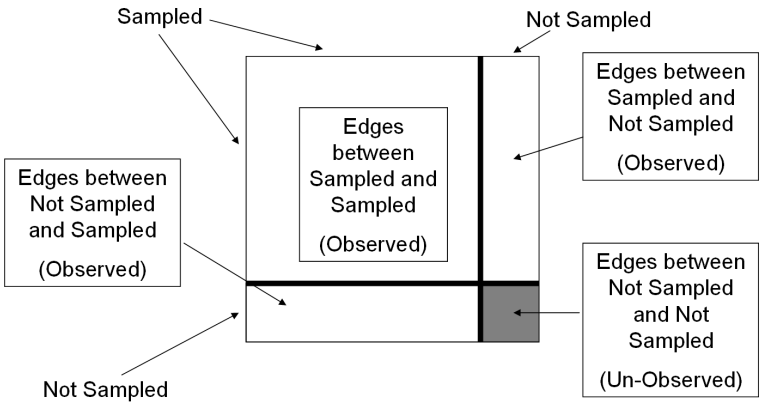

Figure 1: Schematic depiction of sampled and unobserved arc data when the sampling is over an undirected network.

7 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 (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 scale parameter for the GWESP term 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 1. Note that we are taking these parameters as "truth" and considering data produced by sampling from this network.

We conduct all possible datasets produced by a two-wave link tracing starting from two randomly chosen partners (the "seeds"). 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 7. 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 contain the unobserved data, those between the non-sampled.

For each of these samples we use the methods of Section 4 to estimates 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 parameterization and for the mean value parameterization (see Handcock, 2003). waves (corresponding to observing 165 (26%) dyads).

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

There are two isolates and if these are sampled 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 The estimates from these 3 samples are quite variable compared to the other 627 due to the smaller sample size. 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. 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 how far the estimates are apart 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 $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ be alternative parameters for the model (2). The Kullback-Leibler divergence, $KL(\boldsymbol{\xi}, \boldsymbol{\eta})$, of the ERGM with parameter $\boldsymbol{\eta}$ from the ERGM with parameter $\boldsymbol{\xi}$ is:

$$\begin{aligned}
E_{\boldsymbol{\xi}}\left[\log\left(\frac{P_{\boldsymbol{\xi}}(Y=y|X=x)}{P_{\boldsymbol{\eta}}(Y=y|X=x)}\right)\right] &= \sum_{y\in\mathcal{Y}} \log\left(\frac{P_{\boldsymbol{\xi}}(Y=y|X=x)}{P_{\boldsymbol{\eta}}(Y=y|X=x)}\right) P_{\boldsymbol{\xi}}(Y=y|X=x) \\
&= \sum_{y\in\mathcal{Y}} (\boldsymbol{\xi}-\boldsymbol{\eta})\cdot y P_{\boldsymbol{\xi}}(Y=y|X=x) + \log\left(\frac{c(\boldsymbol{\eta})}{c(\boldsymbol{\xi})}\right) \\
&= (\boldsymbol{\xi}-\boldsymbol{\eta})\cdot E_{\boldsymbol{\xi}}[\mathbf{Z}(Y;x)|X=x] + \log\left(\frac{c(\boldsymbol{\eta})}{c(\boldsymbol{\xi})}\right)
\end{aligned}$$

If $\boldsymbol{\xi}$ is the complete data MLE then $E_{\boldsymbol{\xi}}[\mathbf{Z}(Y;x)] = \mathbf{Z}(Y_{obs};x)$ are the observed statistics (given in column 2 of Table 2). The divergence can be easily computed using the MCMC algorithms of Section 6.1.

Figure 7 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 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 parametrizationHandcock (2003).

The properties of the original model's natural parameter estimates are summarized in Table 1. 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 7. A consideration of all samples, leaves the bias small but leads to an increase in the RMSE. The efficiency loss are also substantially increased, especially those of the edges and GWESP terms. For these the errors are, on average, 10% to 21% of the variation in the complete data. However, as we have seem, much of this is from the few extremely bad samples.

Table 2 is the mean value parameterization analog of Table 1. As these are on the same measurement scale as the statistics they are easier to interpret. Again we see they are approximately unbiased when the extreme values are included or excluded. However for these parameters the efficiency loss is small for overall samples.

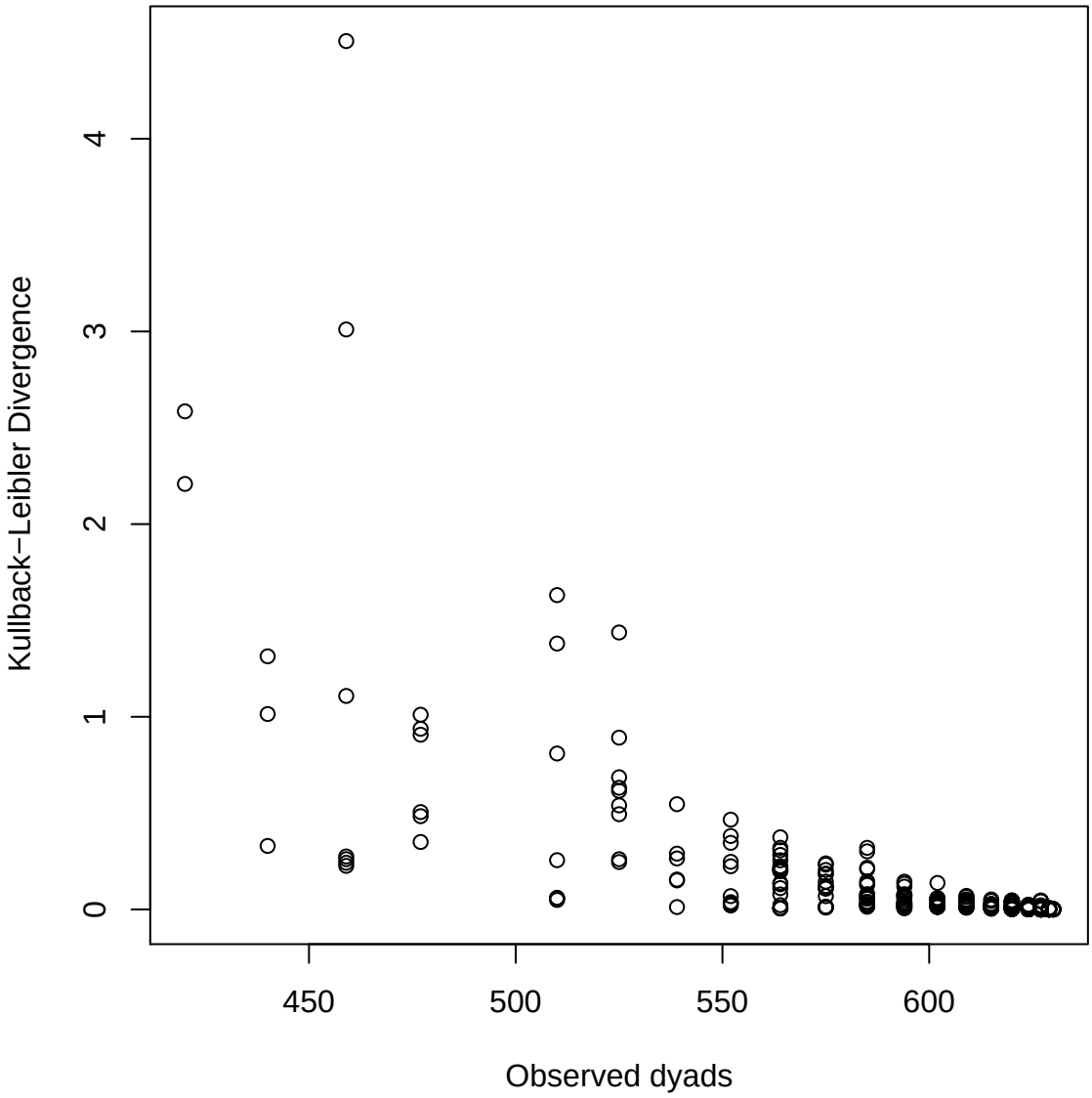

Figure 2: Kullback-Leibler divergence of the MLEs based on the samples compared to the complete data MLE. As the number of dyads sampled increases the information content of the samples approaches that of the complete data. The information loss for the majority of samples is modest.

Table 1: 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 2: 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

8 Application to Adolescent Social Relations

In this section we consider the effect of missing data on friendship nominations in a social network from the National Longitudinal Study of Adolescent Health (Add Health). Add Health is 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 US representative with respect to region of country, urbanicity, school size, school type, and ethnicity (Harris *et. al* 2003). 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 they regarded as their best friends. Thus each

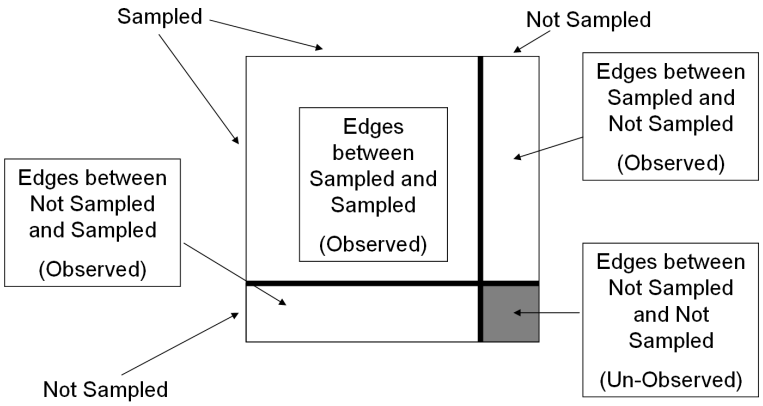

Figure 3: Schematic depiction of observed and unobserved arc data when the missing data is at the respondent level only.

student could nominate up to ten students within the school (Udry 2003).

Here we consider a single school of 89 adolescents from grades seven through twelve. We consider the friendship nominations between all adolescents. The survey also collected the grade and sex of each of the students. The design mechanism of Add Health was a census of the students. However, 19 of the adolescents were not present and did not take the in-school questionnaire. Thus their nominations were missing, although they could be nominated by those who took the survey. Hence of the 7832 nominations $19 \times 88 = 1672$, or 21% were missing. This out-of-design mechanism led to missing observations on the network. In this case there was complete observation of the sex and grade covariates and of the other nominations.

The data pattern is shown in Figure 8. Consider a partition of respondents from non-respondents and the corresponding 2×2 blocking of the sociomatrix, with the four blocks representing arcs from respondents and non-respondents to respondents and non-respondents. The complete data consists of the full sociomatrix. The first two blocks contain the observed data, the arcs sent by respondents, and the second two blocks contain the unobserved data, those sent by non-respondents.

Almost all network analysis of the AddHealth survey models or describes the network among the respondents only, excluding those individuals who are did not complete the survey (Bearman et al., 2004; Harris et al., 2003).

A visualization of the network is given in Figure 4. The non-responding adolescents are colored black and the respondents are colored blue. The student grades are indicated by their text values.

For this study, we fit an ERGM of the form 2 where the set of allowable graphs is restricted to those having no more than five male out-arcs and five female out-arcs for each

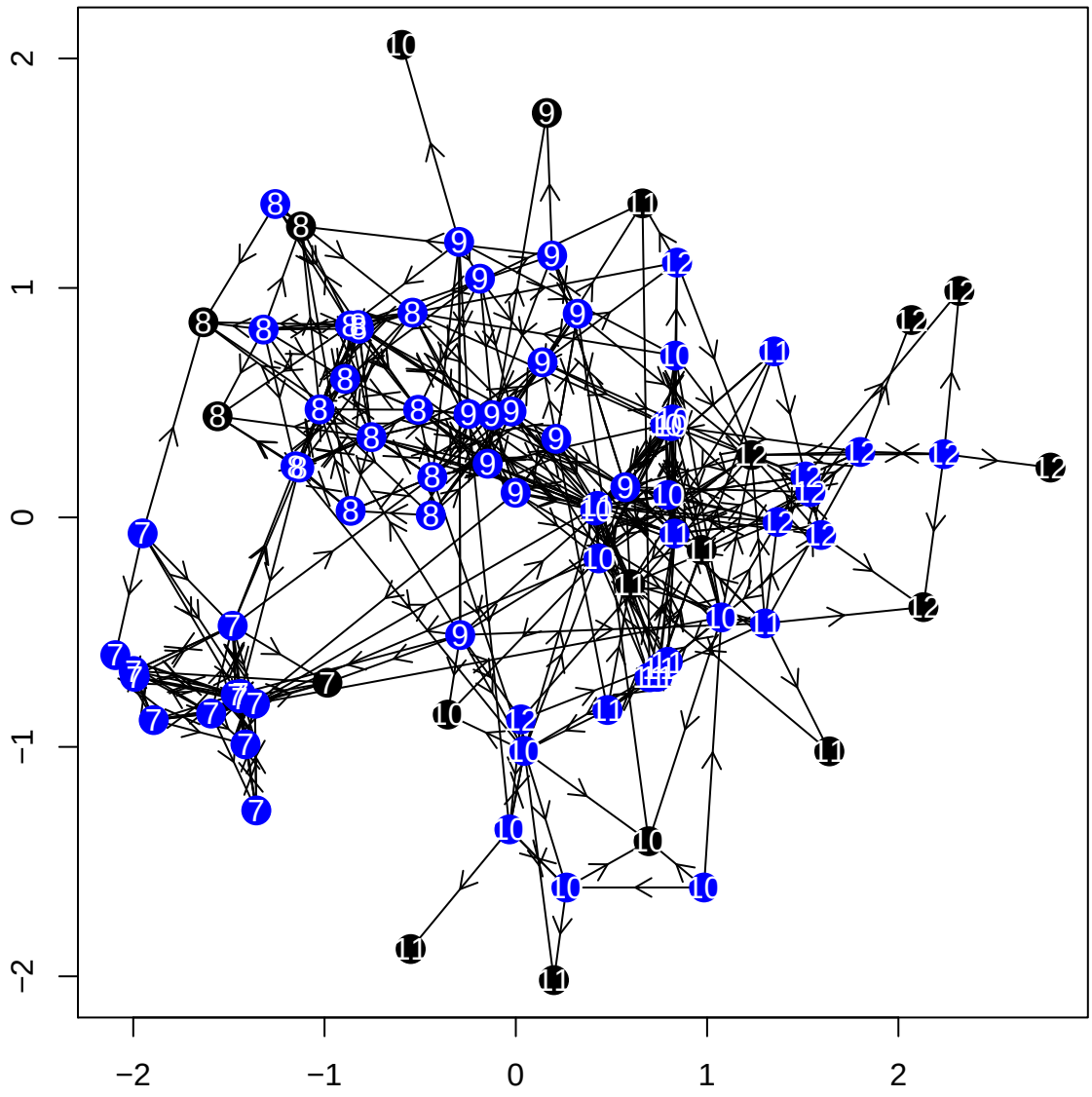

Figure 4: Visualization of the Add Health Network. The non-responding students are colored black and the respondents are colored blue. The arcs represent the presence and direction of the observed nominations. The student grades are indicated by their text values. The layout is using the two-dimensional latent space model of Hoff et al. (2002).

student. This reflects the nature of the relationship modeled and ensures the distribution is over the appropriate sample space.

8.1 Specification of the Model

We specify a model for the social process in which, $\mathbf{Z}(Y; x)$, the set of network statistics has twenty-one terms. This is the same model as in Handcock and Gile (2006), but we repeat the description here for completeness.

The "Density" term represents the overall tendency for students to nominate friends. This captures the overall density of arcs - that is the number of arcs divided by the number of possible arcs in the network.

The "Mutuality" term represents the propensity for arcs to be reciprocated. The corresponding sufficient statistic is the number of dyads with arcs in both directions.

The third through seventh terms capture the differential tendencies for students in different grades to be named as friends. The reference category is 7th grade, so the 8th grade popularity term, for example, captures the degree to which the tendency for 8th graders to receive friendship arcs exceeds that of 7th graders.

The "Male Popularity" term captures the tendency for boys to receive arcs, beyond the tendency of girls.

The following two terms, "girl to same grade boy" and "boy to same grade girl," capture the relative propensity for same-grade arcs to be sent to a same-sex or opposite sex nominee. These effects are allowed to be different for males and females.

The next six terms address the relative tendencies of students to choose friends who are older or younger, closer or farther from their grade, and same or opposite sex. These tendencies are operationalised as dyadic covariates. As an example, consider the "girl to older girl" term. This term applies only to dyads from a female in a lower grade to a female in a higher grade. The coefficient is scaled by the number of years between the two grades. If the "girl to older girl" parameter estimate is α, this means that the log odds of an arc from a girl to a girl one grade older is α more than the log odds of an arc from a girl to a same grade girl. The log odds of an arc to a girl two grades older is an additional α more. Using the number of years older or younger as a linear covariate serves as a way to address the decaying likelihood of friendship across increasing grade differences. The covariates measured here are all measured in terms of "absolute number of years difference."

If these terms were the only ones in the ERGM then each dyad would have arcs independently of every other dyad. Together, they capture the propensities for arcs to be formed between senders and receivers of various characteristics, as well the overall rate at which each group tends to receive arcs. The remaining terms address the dependence or clustering

behaviour among arcs. We know that most arcs are among actors of the same grade and sex, so this is where we focus our attention on the patterns of clustering.

The two triad-based dependence terms address the propensity for arcs to form transitive and cyclical triads respectively among actors of the same grade and sex. These terms are valuable in describing the clustering behaviour in the network. A positive parameter for the transitive triad term suggests that friendships within students of the same sex and grade are likely to form in hierarchal patterns, whereby if Anne nominates Betty (giving greater attractiveness, and greater prestige to Betty), and Betty nominates Carol (even greater prestige), then Anne is more likely to nominate Carol as well. Cyclical triads, on the other hand, can be interpreted as an indication of friendships forming on an egalitarian basis. If Anne nominates Betty (making Anne and Betty close and equal), and Betty nominates Carol, then Carol is more likely to nominate Anne. Together, these two terms capture the flavor of the clustering behaviour in the observed network. In practise, the transitive term is often positive and the cyclical term often negative (as in this case).

The final term "Cyclic Same Grade and Sex" captures one additional facet of clustering. Not all nodes have in arcs. None of the other terms have captured this tendency for some people to simply not be nominated as friends. Therefore, we have included a term to explicitly account for the tendency of the network to contain nodes with no in arcs.

Note that all of these terms measure tendencies with respect to the set of realisations that are possible given the network covariates. The "boy to older boy" term, for example, captures the propensity for arcs from a younger to an older boy, with respect to the total older-younger boy dyads in the network of interest. For this reason, it is reasonable to apply the model fit on the smaller network of respondents only to the larger network of all 89 students.

The model fit reveals many interesting patterns. First, friendship arcs are reciprocated at a higher rate than we would expect at random given the other terms in the model. With regard to grade, 10th graders seem to receive significantly less friendship nominations than the reference 7th graders, although this finding is weaker than the others. There is a complex pattern of sex and grade mixing. In terms of the dyad-dependence terms we see a positive significant transitive triad and a negative significant cyclical triad term. This suggest that friendship arcs within sex and grade tend to form in a hierarchal manner, rather than in an egalitarian regime. This finding is likely the most scientifically interesting of the processes supported by this model. Finally, arcs are clustered so as to produce more nodes receiving no friendship nominations than we would expect from the rest of the terms in the model.

A fuller analysis and description of these results is continued in Handcock and Gile (2006).

	Estimate	s.e.
Density	−1.508	0.19
Mutuality	1.951	0.22
Sex and Grade Factors		
Grade 8 Popularity	−0.171	0.14
Grade 9 Popularity	−0.297	0.16
Grade10 Popularity	−0.346	0.16*
Grade 11 Popularity	−0.052	0.19
Grade 12 Popularity	−0.147	0.18
Male Popularity	0.461	0.16*
Sex and Grade Mixing		
Girl to Same Grade Boy	0.172	0.23
Boy to Same Grade Girl	−0.255	0.23
Girl to Older Girl	−0.928	0.17
Girl to Younger Girl	−1.300	0.22
Girl to Older Boy	−0.882	0.14
Girl to Younger Boy	−1.358	0.23
Boy to Older Boy	−0.859	0.16
Boy to Younger Boy	−1.825	0.35
Boy to Older Girl	−0.641	0.14
Boy to Younger Girl	−1.102	0.19
Transitivity		
Transitive Same Sex and Grade	0.505	0.05
Cyclical Same Sex and Grade	−1.002	0.20
Isolation	3.613	0.68

Table 3: Estimated coefficients and standard errors for the parameters of the model fit adjusting for the missing data pattern.

9 Discussion

In this paper we give a concise and systematic statistical framework for dealing with partially observed network data. The framework includes, but is not restricted to, adaptive network sampling designs and some missing data patterns. We present a definition of an adaptive network design and a result on likelihood-based inference under such designs.

An important simple results 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 indicates that this is not necessary for such samples to be representative. In fact, for the most commonly used designs in practice the designs can be easily taken into account. Hence, despite their form, inference for adaptive" network sampled information is tractable.

We have also shown that, in principle and in practice, it is possible and natural to work with complete graph models even when the data is from a network sample and/or where there is missing data. We have also shown the likelihood framework can naturally adapt to sampling and missing data al la Little and Rubin (2001). This also indicates that the likelihood framework is important to obtain correct inference.

The result that link-tracing designs are adaptive and can be analyzed with likelihood based methods is very valuable in practice and these designs have previously not been able to be analyzed with ERG (or similar) models.

We have also applied the methodology successfully to two different types of unobserved networks.

In our first 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 MLE 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 sample of the same size.

In our second application we show that we can estimate the parameters of a realistic model for a friendship network from the National Longitudinal Study of Adolescent Health. The model also takes into account the fact that the relationship is restricted to those having no more than five male out-arcs and five female out-arcs for each student. This appears to be the first time that either of these two features of friendship data in AddHealth have been adjusted for.

We have made available the code used in this study on the `statnet` website (Handcock et al., 2003).

References

Bearman, P. S., J. Moody, and K. Stovel (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology 110*, 44–91.

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. 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 (2006). 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://csde.washington.edu/statnet.

Harris, K. M., F. Florey, J. Tabor, P. S. Bearman, J. Jones, and R. J. Udry (2003). The national longitudinal of adolescent health: Research design [WWW document]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill, Available at: http://www.cpc.unc.edu/projects/addhealth/design.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002, December). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*(460), 1090–1098.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. with comments by Ronald L. Breiger, Stephen E. Fienberg, Stanley S. Wasserman, Ove Frank and Shelby J. Haberman and a reply by the authors. *Journal of the American Statistical Association 76*(373), 33–65.

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.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 415–444.

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

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.

Udry, R. J. (2003). 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.

White, H. C., S. A. Boorman, and R. L. Breiger (1976). Social-structure from multiple networks I: Blockmodels of roles and positions. *American Journal of Sociology 81*, 730–780.