Estimation in Network Populations

Mike Kwanisai
National Opinion Research Center

Abstract

In social networks subjects are linked to one another forming components and structures that are usually of interest. The subjects may have a variable of interest which, in general, can only be observed after the subjects have been interviewed. Unfortunately, it is sometimes difficult to obtain this information, especially when subjects are rare, hidden or hard-to-reach. Link-tracing sampling designs are commonly used to draw samples from these special populations. Link-tracing designs enrich the sample by following relations from subjects already in the sample to include new subjects. Besides being convenient, these sampling designs produce biased samples that make the estimation of quantities of interest difficult. In this paper we discuss estimation procedures when samples are obtained after following only a fraction of relations. Using simulated and real study data, we demonstrate how estimation for population quantities can be done.

Keywords: Snowball sampling, Adaptive sampling, Link-tracing, Bayesian, Markov Chain Monte Carlo.

1 Introduction

In conventional sampling designs such as simple random sampling, researchers make decisions about the sample size before the sampling procedure begins. In these designs the sampling frame is usually available or can easily be obtained. Obtaining data on the units of interest when conventional sampling designs are used is usually not difficult.

When studying rare, hard-to-reach or evasive human subjects conventional sampling designs do not produce useful data for analysis. For instance, when estimating sensitive human behavioral characteristics such as the use of intravenous (IV) drugs, involvement in commercial sex and any such illegal activities, numerous sampling problems are encountered (Thompson and Collins, 2002). Because of the relative rarity and elusive nature of these populations, conventional sampling designs such as simple random sampling are inefficient for producing data on the individuals of interest.

Human populations however tend to form social networks in which individuals are linked to one another by sociometric relations. In link-tracing sampling designs investigators use these relations or links between subjects to find new subjects to include into the study (Thompson and Frank, 2000; Thompson, 2003; Chow and Thompson, 2003). Any sociometric relation of interest can define a link between two subjects in the population.

Potterat et. al. (1993) used link-tracing designs in a study of a ”high-risk” population in Colorado Springs. The study was on the heterosexual transmission of HIV/AIDS among commercial sex workers and injection drug users (IDUs).

Depending on the manner in which new subjects are included into the sample, different terms have been used for link-tracing designs. Such terms include snowball sampling (Frank, 1979), adaptive sampling (Thompson, 1990, 1992; Thompson and Seber, 1996), respondent-driven sampling (Heckathorn, 1997), random walk sampling (Klovdhal, 1989), chain sampling (Erickson, 1979) and crawling (Burner, 1997). The term web-crawling is used when internet search engines use URL links to sample webpages from the World Wide Web (WWW). One common characteristic of all link-tracing designs is that new subjects are included into the sample by tracing relations or links from subjects already in the sample.

Link-tracing sampling designs conveniently increase the sample size. Link-tracing is sometimes the only easiest and practical way to identify members of rare and hard-to-reach populations (Spreen, 1992; Thompson and Collins, 2002). However, samples obtained using link-tracing designs are biased. This is because subjects with many relationships tend to be over-represented (Spreen, 1992, Kalton and Anderson, 1986; Erickson, 1979, Thompson and Frank 2000). Thus estimation using such samples tend to give biased estimates and getting unbiased estimators is almost impossible without knowledge of subjects' inclusion probabilities.

Depending on the population quantity of interest, samples from network populations can be used to estimate quantities like population size (Frank and Snijders; 1994), network density (Granovetter, 1976), mean degree (Capobianco and Frank, 1982; Frank, 1977b, 1978) or population proportion (Chow and Thompson, 2003; Salganik and Heckathorn, 2004; Thompson and Frank, 2000; Spreen and Coumans, 2000, 2003).

In this paper we focus on estimation of a population proportion when the population size is known.

In section 2 we discuss how samples are obtained from network populations and the estimation methods commonly used. We also introduce the stochastic block model which is the focus of the rest of the paper in section 2. Section 3 discusses model-based estimation methods when samples are obtained after tracing only a fraction of relations. Section 4 show examples and application results using the proposed estimation method. Section 5 concludes the paper with a discussion.

2 Sampling and estimation methods

In link-tracing designs subjects are asked sociometric-type questions and relations with other subjects in the population are obtained. These relations are then traced to include additional subjects into the study. In practice, a small sample of subjects is initially selected at random. These subject are then asked with whom they share special relationship. Newly mentioned subjects, which form the first wave, are then added to the sample. Subjects in the first wave are in turn asked the same question. Also, newly mentioned subjects that are neither in the initial sample nor the first wave are also added to the sample. These subjects form the second wave. The process continues and stops after a certain number of waves, when there are no more newly mentioned subjects or when sample size reaches a predefined size.

Obtaining samples by following all relations is not always practical. For instance large social networks often have problems of high sampling costs and non-responses (Coleman J.S., 1958). This paper will focus on sampling and estimation when some relations are not followed.

2.1 Modelling networks using graphs

Social networks are conveniently modelled when viewed as graphs. Consider a population of size N subjects having relations or links between them. Each subject can be viewed as a node and relations between subjects as edges or links. In most cases each of the N subjects has a value of interest Y.

In a study of injection drug users (IDUs) we can define Y_u as an indicator variable that takes value 1 when subject u is an IDU and 0 otherwise. In general Y can either be a continuous or discrete random variable. A relation between two subjects can be represented by an arc, if it is directional or simply by a link if nondirectional. Thus the population and its relations can be viewed as a graph with node set $V = \{1, \ldots, N\}$ and link set $E = \{(u, v)|u, v \in V\}$.

An adjacency matrix $\mathbf{X}$ is an N by N matrix of node relations with elements $x_{uv} = 1$ if $(u, v) \in E$ or 0 otherwise. For undirected graphs matrix $\mathbf{X}$ is a symmetric matrix of 1's and 0's. If the graph is directional x_{uv} is one if there is an arc from node u to node v. For convenience we may define the diagonal elements of $\mathbf{X}$, that is x_{uu}, to be equal to 0. Thus there are no self links and diagonal elements of the adjacency matrix are zero.

In this paper we only consider undirected relations.

2.2 Estimation methods

In order to make proper inference, the procedure by which the sample is selected must be considered (Thompson and Collins, 2002). Thus, when link-tracing designs are used, computational procedures taking into account the probability of a subject being sampled must be used.

When link-tracing sampling designs are used estimation is complicated because selection or inclusion probabilities cannot be calculated from the sample data for every subject in the sample. This is because for some subjects (or nodes) in the sample investigators do not know how many other subjects would potentially have directed them to the same subject (Frank, 1977). Also, in practice, a well defined probability sampling procedure is usually not used to obtain the initial sample. In a study of IDUs investigators may get the initial sample from any available source like police stations, jails, hospitals and the like.

Two estimation approaches namely design-based and model-based method are generally. Each of these approaches has its advantages and disadvantages.

2.2.1 Design-based estimation

In a design-based estimation approach attempts are made to calculate inclusion probabilities of subjects in the sample without making any assumption made about the population.

Design-based methods have advantages in that they do not assume any model for the population and estimators do not depend on any assumed population model for unbiasedness and consistency. However, the unbiasedness and consistency of such estimators depends on the sampling design carried out.

Design-unbiased estimators have been developed for some link-tracing designs. Design-unbiased estimators are estimators with unbiasedness based on the way the sample is selected but not on any assumptions about the population (Thompson, 1990; Thompson and Seber, 1996). However, design-unbiased estimates can be used with snowball sampling and other graph sampling procedures under certain specific circumstances (Thompson and Collins, 2002). This includes the requirement that all links be followed from individuals in the sample until no new individuals are identified and that there be an initial explicit probability sample.

Because design-based estimators depend on how the sample is obtained, design-based estimation methods cannot be generalized to a wide range of link-tracing designs.

2.2.2 Model-based estimation and the ignorability assumption

Model-based estimation methods assumes that the population of interest follows a certain model. Estimation is then done using the sample data to get population parameter estimates under the model. This is usually achieved by obtaining the observed data likelihood (Thompson and Frank, 2000; Chow and Thompson, 2002).

The main advantage of model-based estimation is that the same method can be applied to a wide range of link-tracing sampling designs (Thompson and Frank, 2000; Thompson and Collins, 2002). Parameter estimation using models has the advantage of making use of well-studied estimation methods such as maximum likelihood and Bayesian methods.

Estimation based on graph sampling is generally difficult to put on a sound statistical basis without assuming a stochastic process giving the original sample or knowledge of the linkage structure for the entire population (Thompson and Collins, 2002; van Meter, 1990).

An assumption commonly used in model-based methods is that of *ignorability*. A link-tracing design is an ignorable one if the probability of obtaining the sample does not depend on the manner in which the initial sample was obtained (Thompson and Frank, 2000).

Consider a population graph with N nodes labelled $1, 2, \ldots, N$ and node values $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$, adjacency matrix $\mathbf{X}$ and a joint probability $f(\mathbf{Y}, \mathbf{X}; \theta, \lambda)$ where θ and λ are parameters of interest. The data d from a sample s, can be expressed as $d = (s, \mathbf{y}_s, \mathbf{x}_s)$ where $\mathbf{y}$ and $\mathbf{x}$ are the realized $\mathbf{Y}$ and $\mathbf{X}$ values. The likelihood can be expressed as

$$L(\theta, \lambda; d) \;\; = \;\; \sum p(s|\mathbf{y}, \mathbf{x}; \theta, \lambda)\mathbf{f}(\mathbf{y}, \mathbf{x}; \theta, \lambda) \qquad (1)$$

where $p(s|\mathbf{y}, \mathbf{x}; \theta, \lambda)$ is the the sampling probability for the sample s and the summation is over all $(\mathbf{y}, \mathbf{x})$ consistent with the data d. If the sample selection depends only on those $\mathbf{y}_s$ and $\mathbf{x}_s$ in the sample, then design probability $p(s|\mathbf{y}, \mathbf{x}; \theta, \lambda)$ can be factored out of equation (1). When this is the case then design and model parameters are distinct and not related (Thompson and Frank, 2000). In such cases the sampling design is said to be ignorable and the likelihood in (1) becomes

$$\begin{aligned} L(\mu, \theta, \lambda; d) \;\; &= \;\; p(s|\mathbf{y}, \mathbf{x}; \mu) \sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} \mathbf{f}(\mathbf{y}, \mathbf{x}; \theta, \lambda) \\ &\propto \;\; \sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \theta, \lambda) \qquad (2) \end{aligned}$$

where μ is the design parameter and the summation is over all unobserved values of $\mathbf{y}$ and $\mathbf{x}$. For ignorable designs, the sampling probability for the sample s, i.e. $p(s|\mathbf{y}, \mathbf{x}; \mu)$, depends only on $(\mathbf{y}, \mathbf{x})$. Thus from (2), the maximum likelihood or Bayes estimators are not affected by the design $p(s|\mathbf{y}, \mathbf{x}; \mu)$.

2.3 A stochastic block model with link probabilities related to node values

Thompson and Frank (2000) proposed a stochastic graph model that assume a population with link probabilities related to node values. For instance, in an HIV/AIDS population we may assume that the probability of a relationship between any two individuals depends on the HIV-status of both individuals.

Consider an undirected graph in which the population size N is known. Let each node u in the graph have a binary variable of interest Y_u such that $P(Y_u = 1) = \theta$. For each pair of nodes u and v define $\lambda_{i+j} = P(x_{uv} = 1|Y_u = i, Y_v = j)$ as the conditional probability of a link between nodes u and v given y values and $i, j = 0, 1$. Note that $\lambda_{i+j} = \lambda_{j+i}$. Also assume that the link indicators (or dyads) x_{uv} are conditionally independent given the node values. This undirected graph model has four parameters of interest $\{\theta, \lambda_0, \lambda_1, \lambda_2\}$.

In a study of IDUs, θ might estimate the probability that a randomly chosen individual from the population is an IDU while λ_0 is the probability of a non-IDU to non-IDU relation.

If M_k is the total number of relations of type k where $k = 0, 1, 2$ and $N_i = \sum_{u=1}^{N}(Y_u = i)$, then the likelihood for the full graph is

$$L(\theta, \lambda; \mathbf{y}, \mathbf{x}) = \prod_{i=0}^{1} \theta_i^{N_i} \prod_{k=0}^{2} \lambda_k^{M_k}(1 - \lambda_k)^{C_k - M_k} \qquad (3)$$

where $k = 0, 1, 2$ and $C_0 = \binom{N_0}{2}$, $C_1 = N_0 N_1$ and $C_2 = \binom{N_1}{2}$. Note that $N = N_0 + N_1$.

Thompson and Frank (2000) gave an expression of the observed data likelihood when the sample is obtained by snowball sampling in which all links are traced, except for the last wave. Chow and Thompson (2003) obtain maximum likelihood and Bayesian estimators for the undirected graph when all links are traced except for the last wave.

The observed data likelihood is complex and difficult to compute when the sample is obtained by tracing only a fraction of relations. In this paper we consider a model-based estimation approach assuming the stochastic block model discussed in this section and ignorability.

3 Estimation when only a fraction of relations is traced

In this section we consider the estimation problem when sample data was obtained after only tracing only a fraction of relations. Three different scenarios arise when samples are obtained in this manner. Firstly it might be possible to uniquely identify the identity of those subjects or nodes at endpoints of the untraced links. For instance, a subject in the sample may know the identity of all his contacts but the investigator, for some reason, traces only a few of relations from the subject. The second case is when nodes at endpoints of untraced links cannot be uniquely identified but, however, known not be outside the observed sample. The third case is when no information at all is known about the identity of endpoints of the untraced links - whether they are in the observed sample or not.

In the following subsections we consider these three cases and assume the model defined in the previous section, together with the ignorability assumption to illustrate how estimation can be done. In this paper we focus only on estimating the population proportion θ.

3.1 Untraced links lead to nodes that can be identified

In this subsection we assume that after tracing only a fraction of links from subjects, untraced links lead to nodes that can be uniquely identified and are outside the observed sample s.

Let $U = (1, 2, \ldots, N)$ be labels for the nodes in the population, s_0 denote the set of nodes for which the value of interest y is observed, $\mathbf{x}_{(s_0,U)}$ denote the set of link indicators within and leading out of nodes in s_0 and s_1 denote the set of nodes in the last wave and obtained by following some of the links from s_0. The y values of nodes in the last wave s_1 are measured but no information about links from s_1 is obtained. It may be that not every link leading out from s_0 is traced. If a link (u, v) is traced from $u \in s_0$, then node v is added to the sample and its y value of interest y_v is observed. If no information is obtained about links out from v, then $v \in s_1$. If the link (u, v) is not traced then the value of y_v is not observed, though the identity of v can be determined. Note that with s_0 information is obtained on node values in it and links leading out from it. Only node values are obtained for s_1 and no links leading out are traced .

When all links are traced except for nodes in last wave s_1, the link indicator matrix $\mathbf{x}_{(s_0,\bar{s})}$ between nodes in s_0 and nodes not in the sample $\bar{s}$ is a zero matrix. When not all links are traced from the sample $s = s_0 \cup s_1$ then $\mathbf{x}_{(s_0,\bar{s})}$ is a matrix of zeros and ones.

If the identity of endpoints v of the untraced links from s_0 is known then the elements x_{uv} of $\mathbf{x}_{(s_0,\bar{s})}$ are 1 when link (u, v) is not traced where $u \in s_0$ and $v \in \bar{s}$. Even when links from s_0 to $\bar{s}$ are known to exist, the y values, $y_{\bar{s}}$ for nodes in $\bar{s}$ are unobserved and unknown. The sample data when untraced links lead to nodes that can be identified is $d = \{(s, \mathbf{y}_s, \mathbf{x}_{(s_0,U)})\}$.

If $\bar{s}$ is the set of nodes that are not in the sample then the unobserved data from the population graph are the y-values in $\mathbf{y}_{\bar{s}}$ and links from and to nodes in the set $s_1 \cup \bar{s}$ which is the sub-matrix $\mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})}$ of the adjacency matrix.

Using data augmentation we can impute the unknown graph quantities $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})}$ and obtain parameter estimates. (Tanner and Wong, 1987; Gilks, 1996; Schafer, 1997).

The conditional distribution of the unobserved node and link-indicator values is

$$P(\mathbf{y}_{\bar{s}}, \mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})}|d) = P(\mathbf{y}_{\bar{s}}|d)P(\mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})}|\mathbf{y}_{\bar{s}}, d) \quad (4)$$

The conditional distribution of y_v, for $v \in \bar{s}$, depends on the y values of nodes in the sample and the link-indicators in the sample that potentially connect to v from $u \in s_0$. Therefore

$$P(\mathbf{y}_{\bar{s}}|d) = \prod_{v\in\bar{s}} P(y_v|y_{s_0}, \mathbf{x}_{(s_0,U)}) \quad (5)$$

Using (5) we can impute the unobserved $\mathbf{y}_{\bar{s}}$. Since the links are unconditionally independent

$$P(\mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})}|y_{\bar{s}}, d) = \prod_{(u,v)} P(x_{uv}|y_u, y_v) \quad (6)$$

Equation (6) can be used to impute unobserved link indicators conditional on the y values.

3.2 Untraced links lead to nodes that cannot be identified but outside the sample

Let us consider the case when untraced links lead to nodes that cannot be identified but are known to be outside the observed sample. The difference with the case discussed previously is that identities of nodes to which untraced links lead are unknown.

Let $\bar{s}$ be the set of nodes not in the sample and s_0^t be the set of nodes in s_0 with all their links traced. Thus all endpoints or y-values of links from nodes in s_0^t are known. Let s_0^u be the set of nodes in s_0 with at least one untraced link. Thus the endpoints of the untraced links from s_0^u are potentially any set of nodes in $\bar{s}$. The set s_0 can therefore be partitioned into two disjoint sets s_0^t and s_0^u where $s_0 = s_0^t \cup s_0^u$. Let $\mathbf{x}_{u+}$ be the vector of the number of untraced links from nodes in s_0^u to nodes in $\bar{s}$. Then the sample data $d = \{(s, \mathbf{y}_s, \mathbf{x}_{(s_0^t,U)}, \mathbf{x}_{(s_0^u,s)}, \mathbf{x}_{u+})\}$.

Although we know the number of untraced links $\mathbf{x}_{u+}$ from s_0^u to $\bar{s}$, link indicators $\mathbf{x}_{(\mathbf{s_0^u},\bar{\mathbf{s}})}$ between nodes in s_0^u and $\bar{s}$ are unknown. Knowing $\mathbf{x}_{(\mathbf{s_0^u},\bar{\mathbf{s}})}$ implies knowledge of all entries of $\mathbf{x}_{(s_0,U)}$ and, as a result, the estimation procedure can use (5) and (6).

A Metropolis-Hastings (MH) algorithm can be used to draw realizations of $\mathbf{x}_{(s_0^u,\bar{s})}$ conditional on $\mathbf{x}_{u+}$ and node values (Hastings, 1970). For every node $r \in s_0^u$, we draw two end nodes a and b from $\bar{s}$ at random and without replacement. Let y_r, y_a and y_b be the y values for nodes r, a and b. The link indicators between node r and a and between node r and b are x_{ra} and x_{rb} respectively. If $x_r^{old} = (x_{ra}, x_{rb})$ and $x_r^{new} = (x_{rb}, x_{ra})$ then conditional on $y_{\bar{s}}$ define acceptance probability α_r as $min\{R_r, 1\}$ where

$$R_r = \frac{P(x_r^{new}|d, y_{\bar{s}})}{P(x_r^{old}|d, y_{\bar{s}})} \quad (7)$$

When $x_{ra} = x_{rb}$ or $y_a = y_b$ then $R_r = 1$. If α_r is greater than a randomly drawn value from $U(0, 1)$, then we accept that the vector of link indicators between node r and a and node r and b is x_r^{new} otherwise it is x_r^{old}. Repeating this draw-switch-accept/reject procedure many times and for all nodes $r \in s_0^u$ give realizations of $\mathbf{x}_{(\mathbf{s_0^u},\bar{\mathbf{s}})}$ which can be used in estimation as described in the previous subsection.

When there are no untraced links $\mathbf{x}_{u+}$, s_0^u is empty and the observed data reduces to $d = \{(s, \mathbf{y}_s, \mathbf{x}_{(s_0,U)})\}$.

3.3 Untraced links lead to any possible set of nodes

In this subsection we consider the case when there is no knowledge as to whether the untraced links would lead to nodes already in the sample s or outside the sample $\bar{s}$. This may be the case in an HIV/AIDS study where interviewed subjects may refuse to give any information that might potentially help to identify their partners.

Let s_0 be the set of nodes in the sample excluding the last wave s_1, s_0^t be the set of nodes in s_0 with all links traced and s_0^u be the set of nodes in s_0 with at least one link untraced. Thus $s_0 = s_0^t \cup s_0^u$. We assume that endpoints of the untraced links from s_0^u lead to nodes either in s or $\bar{s}$. Let $\mathbf{x}_{u+}$ be the vector of the number of untraced links $u \in s_0^u$ and $\mathbf{x}^k_{(s_0^u,s)}$ be the known links from s_0^u to s. The observed sample data is $d = \{(s, \mathbf{y}_s, \mathbf{x}_{(s_0^t,U)}, \mathbf{x}^k_{(s_0^u,s)}, \mathbf{x}_{u+})\}$.

Define $\tilde{s}$ as the set of nodes that can potentially be end points of the untraced links from s_0^u. Note that $\tilde{s}$ include nodes both in s and $\bar{s}$. The unobserved data $(\mathbf{y}_{\bar{s}}, \mathbf{x}_{(s_1\cup\bar{s},s_1\cup\bar{s})})$ can be drawn from the conditional distribution of the unobserved data given sample data d.

For each node $r \in s_0^u$ traced links have known end points while untraced links may lead to $\tilde{s}$, the set of any possible endpoints in the population. Note that $\tilde{s}$ is a subset of $s \cup \bar{s}$.

As described in the the previous subsection, we wish to draw realizations of $\mathbf{x}_{(s_0^u,\bar{s})}$ conditional on $\mathbf{x}_{u+}$ and node values $\mathbf{y} = (\mathbf{y_s}, \mathbf{y_{\bar{s}}})$. For every node $r \in s_0^u$, we draw two nodes a and b at random and without replacement from $\tilde{s}$, the set of nodes that can possibly be endpoints of the untraced links from node r. Defining $x_r^{old} = (x_{ra}, x_{rb})$ and $x_r^{new} = (x_{rb}, x_{ra})$ where x_{ra} and x_{rb} are link indicators between node r and a and between node r and b respectively, the acceptance probability α_r can be calculated using (7). Also α_r is compared to a value randomly drawn from $U(0, 1)$ in order to accept or reject x_r^{new} instead of x_r^{old}. Repeating this process a large number of times for all the $r \in s_0^u$ yields random draws of the unobserved link values $\mathbf{x}_{(s_0^u,\bar{s})}$ given $\mathbf{x}_{u+}$ and $\mathbf{y}$. Considering $\mathbf{x}_{(s_0^u,\bar{s})}$ as known, the estimation process will proceed as discussed in previous sections.

4 Results

In this section we present model-based estimation results using the undirected graph model and approach discussed in the previous sections. Also, we assume that the design is ignorable. We obtained results using data augmentation (Tanner and Wong, 1987; Schafer, 1997) and assuming non-informative priors for the parameter estimates θ and λ's. In this paper we only present result of estimates for θ but the methods discussed here can also be used to obtain estimates for the λ's.

Data augmentation (DA) is an MCMC algorithm that can be used when the complete data is $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} is the observed data and Y_{mis} is the missing part of the data (Tanner and Wong; 1987). The DA algorithm is mainly used to get Bayesian estimates in cases where the complete data posterior $P(\theta|Y_{obs}, Y_{mis})$ is in a simpler form than the observed data posterior (Schafer; 1997). DA has two main steps: namely the imputation (I-) step and the prediction (P-) step. In the I-step we consider a current guess $\theta^{(t)}$ of the parameter and then draw the missing data from the conditional predictive distribution of y_{mis},

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)}). \tag{8}$$

After drawing $Y_{mis}^{(t+1)}$ from $P(Y_{mis}|Y_{obs}, \theta^{(t)})$ the next step is the P-step. In the P-step, we draw the value of θ from its complete data Posterior

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y^{(t+1)}) \tag{9}$$

Repeating steps (8) and (9) after starting with an initial guess $\theta^{(0)}$ yields a sequence $\{Y_{mis}^{(t)}, \theta^{(t)}\}$ with stationary distribution $P(\theta, Y_{mis}|Y_{obs})$. The series $\{Y_{mis}^{(t)}, \theta^{(t)}\}$ defines a Markov chain in which $P(\theta^{(t)}) \rightarrow P(\theta|Y_{obs})$ and $P(Y_{mis}^{(t)}) \rightarrow P(Y_{mis}|Y_{obs})$. Schafer (1997) recommends considering a large "burn in" period t before using the results for estimation because $P(Y_{mis}|Y_{obs}, \theta^{(t)})$ is normally unstable for initial estimates of θ.

4.1 Example 1: A "high-risk" population

Let us consider the Colorado Springs (CS) data from a study on the heterosexual transmission of HIV/AIDS in a "high-risk" population in Colorado Springs (Potterat et. al. 1993). The empirical population had $N = 595$. The y values were 1 for subjects (nodes) who exchanged sex for money and 0 for nodes that did not exchange sex for money. The link-indicator x_{uv} took value 1 if subject u and subject v had a sexual and/or a drug injection relation. This empirical population was also used by Chow and Thompson (2003) as an example in a snowball sampling design in which all links were traced, with the exception of the last wave.

The population quantity of interest is $\theta = 0.223529$. Other quantities that might be of interest for this population, although not considered in this paper, are $\lambda_0 = 0.000855$, $\lambda_1 = 0.004768$, $\lambda_2 = 0.002620$ or average degree = 1.3681.

Table 1 shows estimates of θ when, on average, only a certain percentage of links are traced. For comparison, we compared these results to Bayesian estimates (BE) and Bayesian predictor (BP) values using methods in Chow and Thompson (2003). The estimates for θ were comparable with results from link-tracing only a fraction of relations giving smaller mean square errors. These results were obtained after 5000 iterations.

Although these results are encouraging, they do depend on a number of factors such as model fitness, ignorability assumption, population size and the quantity being estimated.

4.2 Example 2: A hypothetical population

In this example we consider a hypothetical population of size $N = 100$ and an average degree of 7.5. The proportion of 1's in the population is 0.29.

Table 2 shows estimates of θ when 90%, 80% and 60% of links are traced. We consider the case when untraced links lead to endpoints that are known and when untraced links lead to unknown endpoints.

For both the two cases, estimates for θ are comparable to Bayesian estimates (BE) and Bayesian predictors (BP). Mean square error values are smaller when untraced links lead to nodes that can be identified. As seen on Table 2, the naive estimator overestimates θ and gives the largest mean square error. These results were obtained after 5000 iterations.

5 Discussion

Link-tracing designs are useful in gathering information on drug use, sexual behavior or such similar activities from hard-to-reach and elusive populations. Several link-tracing designs have been developed and their use depends mostly on the focus of the study and the researcher's choice. It is usually hard for the researcher to decide the best stopping criteria once the link-tracing has started. For instance, in snowball sampling the researcher might choose to stop sampling new subjects when all subjects with known relations are in the sample (Frank and Snijders, 1994) or after sampling a certain number of waves. The stopping criteria may also depend on sample size especially if there are resource limitations to trace more links.

Estimation based on data gathered by a link-tracing design is an area of active ongoing research. Depending on the estimation procedure used, it is sometimes very hard to obtain unbiased estimators especially when many waves are considered and not all links are traced. This problem is compounded when a design-based estimation approach is used because computing inclusion probabilities gets harder as more waves are considered.

Model-based estimation has the advantage of being generalizable and that it can be done with the aid of well developed MCMC methods.

Design-based estimation rely on being able to compute inclusion probabilities. The inclusion probability for subject i is the probability π_i that subject i is included into the sample. Inclusion probability is hard to compute without knowledge

Table 1: Estimates of population proportion using the Colorado Springs ”high-risk” data. The total population size is N=595 and true population proportion $\theta = 0.2235$. An initial sample of 40 was link-traced until a sample of size 85 was observed. These results are for the case where endpoints of the untraced links are known. These results were obtained after 5000 iterations.

	All links traced		*Not all links traced*		
	BE	*BP*	*90% traced*	*80% traced*	*70% traced*
Avg waves	*1*	*1*	*1.63*	*1.74*	*1.82*
Estimate	0.2130	0.2093	0.2074	0.2073	0.2008
MSE	0.0029	0.0031	0.0027	0.0027	0.0025

Table 2: Estimates of the proportion of 1’s in a hypothetical population of size 100 and average degree 7.5. The true proportion was $\theta = 0.29$. An initial sample of 5 was link-traced until a sample size of 35 was sampled. These results were obtained after 5000 iterations.

				Endpoints known			*Endpoints unknown*		
	All links traced			*% of links traced*			*% of links traced*		
	Naive	*BE*	*BP*	*90%*	*80%*	*60%*	*90%*	*80%*	*60%*
Avg waves	*1*	*1*	*1*	*1.00*	*1.01*	*1.14*	*1.00*	*1.03*	*1.28*
Estimate	0.4329	0.3008	0.2968	0.2997	0.3000	0.2895	0.3106	0.3012	0.3022
MSE	0.0350	0.0024	0.0025	0.0013	0.0013	0.0012	0.0035	0.0036	0.0028

of the entire population network. Sometimes approximations are used to estimate π_i. Heckathorn (2001) used the argument that in one-wave designs the inclusion probability for a node i is proportional to its degree d_i to derive estimators for a respondent-driven sampling study. In most practical cases one needs to know the population size, which is usually unknown in hidden populations.

Mathematical models for human populations are complex and solving them to get estimates is generally difficult (Newman et. al., 2002; Thompson and Frank, 2000). Also, sometimes assumptions used in population models, although mathematically convenient, they tend to be unrealistic, especially when human populations are involved. Assuming a model especially for human populations is a topic of fierce debate among researchers. Even when models are suggested, in most cases they are complex and hard to implement. Holland and Leinhardt (1981) discussed an exponential family of models, now commonly known as p^* models (Andersen et. al, 1999; Wasserman and Pattison, 1996).

In this paper we discussed model-based estimation procedures and applied a simple stochastic block model to data obtained from tracing only a fraction of relations. Using the model and an MCMC method, we illustrated how estimation can be done and generalized. Although model-based estimation is computationally appealing, it is hard to imagine a good model for human populations because human behavior tends to be complex.

Acknowledgements

The author wishes to thank Professor Steven K. Thompson for his priceless advice, S. Q. Muth for providing the data used in this paper and Susan Hinkins for her constructive comments and support during the course of writing this paper.

References

Capobianco, M. and Frank, O. (1982), “Comparison of statistical graph-size estimators,” *Journal of Satistical Planning and Inference*, **6**, 87-97.

Anderson, C.J., Wasserman, S., and Crouch, B. (1999), “A P* Primer: Logit models for social networks,” *Social Networks*, **21**, 37-66.

Burner, M. (1997), “Crawling towards eternity: Building an archive of world wide web,” *Web Techniques Magazine*, **2**(5), May 1997

Chow, M. and Thompson, S.K. (2003), “Estimation with link-tracing sampling designs - A Bayesian approach,” *Survey Methodology*, **29**(2), 97-205

Coleman, J.S. (1958), “Snowball Sampling: Problems and techniques of chain referral sampling,” *Human Organization*, **17**, 28-36.

Erickson, B.H. (1979), “Some problems of inference from chain data,” *Sociological Methodology*, **10**, 276-302.

Frank, O. (1977), “Survey sampling in graphs,” *Journal of Statistical Planning and Inference*, 235-264.

Frank, O. (1978), “Estimation of the number of connected components in a graph by using a sampled subgraph,” *The Scandinavian Journal of Statistics* **5**, 177-188.

Frank,O. (1977b), “Estimation of Graph Totals,” *Scandinavian Journal of Statistics* **4**, 81-89.

Frank,O. (1978), “Sampling and estimation in large social networks,” *Social Networks*, **1**, 91-101.

Frank, O. (1979), “Estimation of population totals by use of snowball samples,” *In Perspectives on Social Network Research*, Ed. P. Holland and S. Leinhardt, New York: Academic press, 319-347.

Frank, O. (1988), “Random sampling and Social networks: a survey of various approaches,” *Mathmatiques, Informatique et Sciences humaines*, **26**, 19-33.

Frank, O. and Snijders, T. (1994), "Estimating the size of hidden populations using snowball sampling," *Journal of Official Statistics*, **10**, 53-67.

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

Gilks,W.R., Richardson, S., and Spiegelalter, D.J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall

Granovetter, M. (1976), "Network sampling: some first steps," *American Journal of Sociology*, **81**, 1287-1303.

Hastings, W.K. (1970), "Monte Carlo sampling using Markov Chains and their applications," *Biometrika*, **57**, vol. 1, 97-109.

Heckathorn, D. D. (1997), "Respondent-driven sampling: a new approach to the study of hidden populations," *Social Problems*, **44**, 174-199.

Holland, P.W. and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, **76**, 33-65.

Kalton, G. and Anderson, D. W. (1986), "Sampling rare populations," *J. R. Stat. Soc.* Ser. A ,**149**, 65-82.

Klovdahl, A (1989), "Urban social networks: some methodological problems and possibilities," *In: Kochen, M. (Ed), The small world.* Ablex Publishing, Norwood, NJ 176-210.

Kwanisai, M. (2004), "Estimation in link-tracing designs with subsampling," Ph.D. Thesis, The Pennsylvania State University.

Newman, M.E., Watts, D.J. and Strogatz, S.H. (2002), "Random graph models for social networks," *Proceedings of the National Academy of Sciences of the United States of America*, **99**, Suppl. 1, 2566-2572.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993), "AIDS in Colorado Springs: Is there an epidemic?," *AIDS* **7**, 1517-1521.

Rubin, D.W. (1976), "Inference and missing data," *Biometrika*, **63**, 581-592.

Salganik, M.J. and Heckathorn, D.D. (2004), "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, **34**, 193-239.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Spreen, M. (1992), "Rare populations, hidden populations, and link-tracing designs; what and why?," *Bulletin de Methodologie Sociologique*, **36**, 34-58.

Spreen, M., Coumans. M. (2003), "A note on network sampling in drug abuse research", *Connections*, **25** (1): 27-35

Spreen, M., Coumans. M. (2000), "Network sampling hard drug users. A structural analysis of the clients of aid agencies Heerlen," *To appear: Kwantitatieve Methoden*

Tanner, W.A. and Wong, W.H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the Royal Statistical Society*, Series B, **62**, 795-809.

Thompson, S.K. (1990), "Adaptive cluster sampling," *Journal of the American Statistical Association*, **85**, 1050-1059.

Thompson, S.K. (1992), *Sampling*, New York: Wiley.

Thompson, S.K. (1991), "Stratified adaptive cluster sampling," *Biometrika*, **78**, 389-397.

Thompson, S.K. (2003), "Markov Chain Sampling Designs in Graphs," *Technical report*, **03-01**, Dept. of Statistics, The Pennsylvania State University.

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

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

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

van Meter, K.M. (1990), "Methodological and design issues: techniques for assessing the representatives of snowball samples," *In: Lambert, E.Y. (Ed.), The Collection and Interpretation of Data from Hidden Populations*, National Institute on Drug Abuse **898**, 31-43.

Wasserman, S. and Pattison, P. (1996), *Social Network Anaylsis: Methods and Applications*, New York and Cambridge: Cambridge University Press.