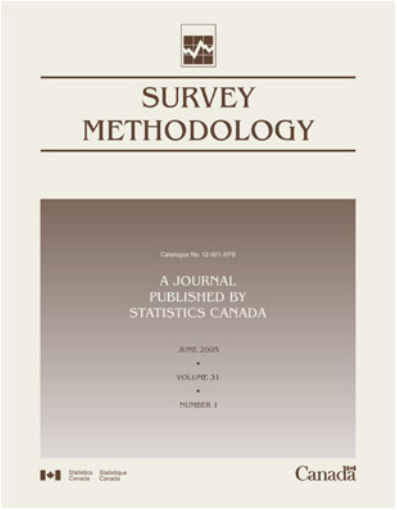
Catalogue no. 12-001-XIE

Survey Methodology

December 2003



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2003

Published by authority of the Minister responsible for Statistics Canada

February 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 2, pp. 197-205

Estimation with Link – Tracing Sampling Designs – A Bayesian Approach

Mosuk Chow and Steven K. Thompson [1]

Abstract

In link-tracing designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. In this paper, we propose a Bayesian approach for the estimation problem. For studies using link-tracing designs, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield better estimators. When the available information is vague, one can use noninformative priors and conduct a sensitivity analysis. In our example we found that the estimators were not sensitive to the specified priors. It is important to note that, under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added difficulty. By contrast, such tasks are difficult to perform using the classical approach. In general, a Bayesian analysis yields one distribution (the posterior distribution) for the unknown parameters, and from this a vast number of questions can be answered simultaneously.

Key Words: Link-tracing designs; Snowball samples; Adaptive sampling; Graph sampling; Network sampling; Beta prior.

1. Introduction

Social network data include measurements on the relationships between people or other social entities as well as measurements on entities themselves. Collecting network data on entire networks requires a great deal of time and effort, especially when networks are large. It is thus important to be able to estimate network properties from samples. In link-tracing sampling designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. For example, in a study of injection drug use in relation to the spread of the HIV infection, social leads from initial respondents may be traced and the linked individuals added to the sample. (*e.g.*, see Neaigus, Friedman, Goldstein, Ildefonso, Curtis and Jose 1995; Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais 1996 and Thompson and Collins 2002). Similarly, for studies of homeless people, respondents may be asked about other homeless people who will then be sampled.

Populations with social structure are often modeled as graphs, with the nodes of the graph representing populations and the arcs of the graph representing social links, relationships, or transactions. In the graph setting, the variables of interest include both those associated with nodes and those associated with pairs of nodes. The population graph itself can be viewed either as a fixed structure or as a realization of a stochastic graph model. Samples are taken to obtain information about the population graph. Usually, the sampling method will take advantage of the arcs or links from one entity to another.

There is a large literature on network sampling, both applied and theoretical. Frank (1977a, 1977b, 1977c, 1978, 1979, 1980, 1997) has many important results in sampling for social networks. His classic work (Frank 1971) presents basic solutions for estimating graph quantities from the sample data. Snijders and Nowicki (1997) propose various statistical approaches, including a Bayesian approach, for estimation and prediction with stochastic blockmodels for graphs in which the node values are not observed.

Snowball sampling (Goodman 1961) is one type of link-tracking sampling design in which individuals in an initial sample are asked to identify acquaintances, who in turn were asked to identify acquaintances, and so on for a fixed number of stages or waves. Erickson (1978) and Frank (1979) review snowball sampling designs with the goal of understanding how other "chain methods" (methods designed to trace ties through a network from a source to an end) can be used in practice. Snijders (1992) used the same term "snowball sampling" to include designs in which only a subsample of links from each node is traced. Frank and Snijders (1994) consider model and design-based estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Another link-tracing procedure for which design-based estimators are available is adaptive cluster sampling Thompson and Seber 1996), which has been formulated in the graph setting as well as the spatial setting.

With a fixed-population, design-based approach in the graph setting, both the characteristics of the people and the social network structure of the population are viewed as fixed, unknown values. The properties such as design-unbiasedness do not depend on any assumption about the

1. Mosuk Chow and Steven K. Thompson, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A.

population itself but they do depend on the sampling design being carried out as specified. In this paper, we consider the model-based methods since they can be applied to a wide range of sample selection procedures. In many studies of hidden and hard-to-reach populations, the sample selection procedures, including link-tracing, are not readily analyzed based on idealized design induced probabilities, but results from the model-based methods can be applied for the cases.

Thompson and Frank (2000) used a model-based approach to inference with link-tracing designs. In their paper, maximum likelihood estimators of population graph parameters and predictors of realized population graph quantities were described. In this paper, we adopt a Bayesian approach fro the graph estimation problem. For real problems with sampling designs that follow social links from one person to another, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield improved estimators. Moreover, when the available information is vague, we can use noninformative priors and conduct a sensitivity analysis. It is important to note that under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added difficulty whereas such tasks would be difficult to perform using the maximum likelihood approach. We deal with inferences for both the characteristic of nodes and also of arcs such as the prevalence of disease in a certain community and also the transmission rate of that disease between two subjects.

Notation for a full graph model with links related to node values and its likelihood function will be given in section 2. In section 3, the likelihood function for the simple obtained from a link-tracing design will be presented and a Bayesian inference method will be introduced. In section 4, an illustrative example will be given. The paper will be concluded by an empirical example and a discussion in section 5.

2. The Model

Using notation similar to Frank (1971) and Thompson and Frank (2000), we denote the full set of node labels by $U = \{1, 2, ..., N\}$ which from the population of N units. A variable of interest associated with an individual node u will be denoted Y_u while a variable of interest associated with pair of nodes u and v will be denoted A_{uv}. The sequence of node variables of interest is denoted by $\mathbf{Y} = (Y_1, ..., Y_N)$. Here we consider the variable of interest A_{uv} as an indicator variable which equals one if there is an arc (directional link) from u to v and zero otherwise for two distinct nodes u and v. The matrix of arc indicators, having A_{uv} as the element in the u^{th} row and v^{th} column, is the graph adjacency matrix, denoted $\mathbf{A}$. For convenience we will assume that the diagonal elements A_{uu} are zero. The ordered pair (u, v) is referred to as a dyad of type $(Y_u, Y_v; A_{uv}, A_{vu})$. In the following assumed model the node variables $Y_1, ..., Y_N$ are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0.1$, and $\theta_0 + \theta_1 = 1$. Conditional on the node values $Y_1, ..., Y_N$, the dyads (A_{uv}, A_{vu}) are independent, for $1 \le u < v \le N$, with conditional distribution given by $P[(A_{uv}, A_{vu}) = (k, l) \mid Y_u = i, Y_v = j] = \lambda_{ijkl}$ for all combinations of $i = 0.1$; $j = 0.1$; $k = 0.1$; and $l = 0.1$. For all combinations of i and j, the sums over k and l are denoted $\lambda_{ij..} = \sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order to get graph probabilities not depending on node identities, the following natural symmetry conditions are assumed: $\lambda_{1110} = \lambda_{1101}$, $\lambda_{1011} = \lambda_{0111}, \lambda_{1010} = \lambda_{0101}, \lambda_{1001} = \lambda_{0110}, \lambda_{0010} = \lambda_{0001}$ and $\lambda_{1000} = \lambda_{0100}$. For example, the first and the fifth conditions say that between two nodes having the same value, the probability of an arc in either direction is the same. Let N_i denote the total number of nodes with value i in the graph so that $N_0 + N_1 = N$. Let further M_{ijkl} denote the total number of dyads of type $(ijkl)$, that is, the total number of ordered node pairs (u, v) such that $(Y_u, Y_v; A_{uv}, A_{vu}) = (ijkl)$. The likelihood for the full graph under the model with parameters (θ, λ) is $L(\theta, \lambda; Y, A) = (\prod_{i=0}^{1} \theta_i^{N_i})(\prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{k=0}^{1}\prod_{l=0}^{1}\lambda_{ijkl}^{M_{ijkl}})$.

3. Bayesian Inference from Link-Tracing Designs

3.1 Likelihood Function given the Sample Data

A sample s from the graph is a subset of nodes from U and a subset of node pairs from U^2. The sample data $d = (s, y_s, a_s)$ are a function of the sample selected and of the graph values y and a. For any designs in which the selection of the sample depends on graph y and a values only through those values y_s and a_s included in the data, the design does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators (Rubin 1976, Thompson and Frank 2000). For example, many of the snowball and other link-tracing designs are ignorable for likelihood-based inference provided the selection procedure for the initial sample is ignorable. Any carefully implemented conventional or adaptive survey design would be ignorable in this sense. Nonignorable initial samples can occur when the selection is uncontrolled and selection probabilities are related to unobserved node and link values, as when people with risk-averse behaviors and low numbers of relationships are less conspicuous to investigators, thereby influencing what units are missed and hence influencing sample selection probabilities in ways that are not measured.

Consider the link-tracing design in which an initial sample s_0 is selected and all links out from nodes in s_0 are followed to add the set s_1 of nodes not in s_0 that are adjacent to nodes in s_0. The whole sample is $s = s_0 \cup s_1$. The entire set of labels in the population can be written as

the union of three disjoint sets, $U = s_0 \cup s_1 \cup \bar{s}$ where $\bar{s}$ denotes the nonsampled nodes. Here, we consider a design in which the decision to follow the links from node u depends on the node value y_u. For example, in a study on injection drug use, the initial sample may contain both users and nonusers. If the investigators choose to follow social links only from users, then the design depends adaptively on the node $y-$values as well as the links. The design then can be written $P(s \mid y_s, a_{s_0 U})$, since the selection procedure depends on both node and link values. The data are $d = (s, y_s, a_{s_0 U})$. Since the decision depends on y and a values only through the observed data, the design factors out of the likelihood function and divides out of the Bayes posterior, so that likelihood or Bayes inference depends only on the assumed model.

With the graph model described in the previous section, it then follows (Thompson and Frank 2000) that the likelihood with the sample data is:

$$L(\theta, \lambda; d) = P(s \mid y_s, a_{s_0 U}) \sum \left(\prod_{u=1}^{N} \theta_{y_u} \right) \left(\prod_{u<v} \lambda_{y_u y_v a_{uv}, a_{vu}} \right)$$

where the sum is over all values of y_u and a_{uv} that are not fixed by the sample data.

For link-tracing designs in which all links, rather than a subsamplе, from the initial sample nodes are traced, all of the elements in the submatirx $a_{s_0 \bar{s}}$ are zero. It has been shown by Thompson and Frank (2000) that the likelihood function can then be written as:

$$L(\theta, \lambda; Y, A) = P(s \mid y_s, a_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \left(\prod_{ijk} \lambda_{ijk.}^{m_{ijk.}(s_0, s_1)} \right) \times \left[\sum_j \theta_j \prod_i \lambda_{ij0.}^{n_i(s_0)} \right]^{n(\bar{s})} \quad (1)$$

where $n_i(s)$, $n_i(s_0)$, and $n_i(\bar{s})$ denote the numbers of nodes of type i in the full sample s, the initial sample s_0, and the nonsampled nodes $\bar{s}$, respectively, and $m_{ijkl}(s_0, s_0)$, $m_{ijkl}(s_0, s_1)$ are the counts of node pairs in $s_0 \times s_0$ and $s_0 \times s_1$.

For a symmetric model, $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always two-way or, equivalently, they can be considered as undirected edges. The full symmetric model has parameters λ_{ijkk} λ_{jikk} for $i, j, k = 0.1$, with $\lambda_{ij00} + \lambda_{ij11} = 1$. To simplify notation for this model, let $\beta_{i+j} = \lambda_{ij11}$ and thus β_k denotes the probability of a mutual link between two nodes having total value k, for $k = 0.1$ or 2. The above likelihood simplifies to

$$L(\theta, \beta; d) = P(s \mid y_s, a_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{i,j} \beta_{i+j}^{m_{ij11}(s_0, s)} (1-\beta_{i+j})^{m_{ij00}(s_0, s)} \right) \times \left[\sum_j \theta_j \prod_i (1-\beta_{i+j})^{n_i(s_0)} \right]^{n(\bar{s})} . \quad (2)$$

Now define $r_{0,0} = m_{0000}(s_0, s)$, $r_{0,2} = m_{0011}(s_0, s)$, $r_{1,0} = m_{0100}(s_0, s) + m_{1000}(s_0, s)$, $r_{1,2} = m_{0111}(s_0, s) + m_{1011}(s_0, s)$, $r_{2,0} = m_{1100}(s_0, s)$, $r_{2,2} = m_{1111}(s_0, s)$. Note that the r's are dyad counts where the first index represents the sum of the node values and the second index represents the sum of the link values. The above expression can be rewritten as:

$$L(\theta, \beta; d) = P(s \mid y_s, a_{s_0 U}) \theta_0^{n_0(s)} (1-\theta_0)^{n_1(s)} \beta_0^{r_{0,2}} (1-\beta_0)^{r_{0,0}} \beta_1^{r_{1,2}} (1-\beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1-\beta_2)^{r_{2,0}} \left[\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} (1-\beta_2)^{n_1(s_0)} \right]^{n(\bar{s})} . \quad (3)$$

In the remainder of this paper, we focus on the full symmetric model to illustrate the proposed Bayesian methodology for simplicity of presentation. The same method can be applied to the general model with the likelihood function given in (1).

3.2 Choice of Prior Distributions

Since there are no specific constraints on θ_0, β_0, β_1, β_2, we may assume independent priors on θ_0, β_0, β_1, β_2, all of which take values in the interval $[0, 1]$. It is quite common to put a beta prior on a parameter that takes values in $[0, 1]$ because most smooth unimodal distributions on $[0, 1]$ can be well approximated by some beta distributions and the class of beta distributions is reasonably rich to model the uncertainty about the parameter. Also, the expression in (3) is in general quite complex but beta priors can yield a tractable posterior distribution (to be shown later). Using beta priors, we obtain an analytic formula for the Bayes estimates and the marginal posterior distribution.

In this paper we consider independent beta priors for the parameters:

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1} (1-\theta_0)^{b-1} \beta_0^{c-1} (1-\beta_0)^{d-1} \beta_1^{e-1} (1-\beta_1)^{f-1} \beta_2^{g-1} (1-\beta_2)^{h-1} . \quad (4)$$

When determining the constraints a and b it is often useful to equate the mean $E[\theta_0] = a/(a+b)$ of $\text{Beta}(a, b)$ to a value which represents your belief about the location of θ_0 and the variance $\text{Var}[\theta_0] = ab/(a+b)^2 (a+b+1)$ of $\text{Beta}(a, b)$ to a value which represents the uncertainty put on the specified θ_0 value. Similarly, the values of c, d, e, f, g and

h can be determined. For example, if one is interested in the prevalence of injection drug use in a certain community, one may take an initial sample and trace links by asking the injection drug user in the sample to name the people with whom they share injection equipment. If the value $y_u = 1$ represents injection drug use, then θ_0 is the percentage of non-users in that community. Quite often an estimate for the central location and the spread of θ_0 may be provided.

In the case of complete ignorance, we will consider three commonly used noninformative priors and provide a comparison of the resulting Bayes estimates in our illustrative example in section 4. (For a fuller discussion of the noninformative priors, see Berger 1985, pages 89–90). The first one is the uniform prior, which corresponds to $\text{Beta}(1, 1)$. The second one, $\text{Beta}(0, 0)$, suggested by Haldane (1931), has an improper density. It is equivalent to a prior uniform in the log-odds $\log\{\theta_0/(1-\theta_0)\}$. A possible compromise between $\text{Beta}(1, 1)$ and $\text{Beta}(0, 0)$ is $\text{Beta}(1/2, 1/2)$, which has a proper density. This prior implies a uniform prior for $\sin^{-1}\sqrt{\theta_0}$.

3.3 Posterior Distribution and Bayes estimates

In our problem, the posterior distribution $\pi(\theta_0, \beta_0, \beta_1, \beta_2 \mid d)$ corresponding to the beta priors is given by:

$$\begin{aligned}\pi(\theta_0, \beta_0, \beta_1, \beta_2 \mid d) \propto\ & \theta_0^{n_0(s)+a-1}(1-\theta_0)^{n_1(s)+b-1} \\ & \beta_0^{r_{0,2}+c-1}(1-\beta_0)^{r_{0,0}+d-1} \\ & \beta_1^{r_{1,2}+e-1}(1-\beta_1)^{r_{1,0}+f-1} \\ & \beta_2^{r_{2,2}+g-1}(1-\beta_2)^{r_{2,0}+h-1} \\ & \left[\theta_0(1-\beta_0)^{n_0(s_0)}(1-\beta_1)^{n_1(s_0)} \right. \\ & \left. \quad +(1-\theta_0)(1-\beta_1)^{n_0(s_0)}(1-\beta_2)^{n_1(s_0)}\right]^{n(\bar{s})}. \end{aligned} \quad (5)$$

To find the posterior mean (Bayes estimate) of θ_0, let

$$\begin{aligned} q(\theta_0, \beta_0, \beta_1, \beta_2) =\ & \theta_0^{n_0(s)+a-1}(1-\theta_0)^{n_1(s)+b-1} \\ & \beta_0^{r_{0,2}+c-1}(1-\beta_0)^{r_{0,0}+d-1} \\ & \beta_1^{r_{1,2}+e-1}(1-\beta_1)^{r_{1,0}+f-1} \\ & \beta_2^{r_{2,2}+g-1}(1-\beta_2)^{r_{2,0}+h-1} \\ & \left[\theta_0(1-\beta_0)^{n_0(s_0)}(1-\beta_1)^{n_1(s_0)} \right. \\ & \left. +(1-\theta_0)(1-\beta_1)^{n_0(s_0)}(1-\beta_2)^{n_1(s_0)}\right]^{n(\bar{s})}. \end{aligned}$$

Since $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = B(\alpha, \beta)$ is the beta function, we have the following two results:

$$\begin{aligned} M_1 =& \int_0^1\int_0^1\int_0^1\int_0^1 q(\theta_0, \beta_0, \beta_1, \beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2 \\ =& \sum_{i=0}^{n(\bar{s})}\binom{n(\bar{s})}{i} B(n_0(s)+a+i,\ n(\bar{s})+n_1(s)+b-i) \\ & B(r_{0,2}+c, i\ n_0(s_0)+r_{0,0}+d)\,B(r_{1,2}+e, i\ n_1(s_0) \\ & +(n(\bar{s})-i)n_0(s_0)+r_{1,0}+f) \\ & B(r_{2,2}+g,\ (n(\bar{s})-i)n_1(s_0)+r_{2,0}+h). \end{aligned}$$

$$\begin{aligned} M_2 =& \int_0^1\int_0^1\int_0^1\int_0^1 \theta_0 q(\theta_0, \beta_0, \beta_1, \beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2 \\ =& \sum_{i=0}^{n(\bar{s})}\binom{n(\bar{s})}{i} B(n_0(s)+a \\ & +1+i,\ n(\bar{s})+n_1(s)+b-i) \\ & B(r_{0,2}+c,\ i\ n_0(s_0)+r_{0,0}+d) \\ & B(r_{1,2}+e, i\ n_1(s_0)+(n(\bar{s})-i)n_0(s_0)+r_{1,0}+f) \\ & B(r_{2,2}+g,\ (n(\bar{s})-i)n_1(s_0)+r_{2,0}+h). \end{aligned}$$

The Bayes estimate for θ_0 can thus be evaluated by the quotient of the righthand side of the above two equations since:

$$\begin{aligned} E(\theta_0|d) &= \frac{\int_0^1\int_0^1\int_0^1\int_0^1 \theta_0 q(\theta_0, \beta_0, \beta_1, \beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2}{\int_0^1\int_0^1\int_0^1\int_0^1 q(\theta_0, \beta_0, \beta_1, \beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2} \\ &= \frac{M_2}{M_1}. \end{aligned}$$

Similarly, the Bayes estimates for β_0, β_1, β_2 can be computed.

3.4 Prediction of Realized Graph Quantities

Consider the problem of estimating or predicting, from the sample data, the realized value of some graph quantity $Z = Z(\mathbf{Y}, \mathbf{A})$, an observable but unobserved finite-population quantity. Denoting the unknown parameters collectively by ψ, the relevant posterior predictive density is

$$\begin{aligned} f(z \mid d) &= \int f(z \mid d, \psi)\pi(\psi \mid d)\,d\psi \\ &\propto \int f(d, z \mid \psi)\pi(\psi)\,d\psi \end{aligned} \quad (6)$$

where the constant of proportionality is, as usual, $f(d)$.

For example, suppose the objective is to predict the proportion of nodes in the population that have value $y = 1$. Let $n_1(s)$ denote the number of nodes for which $y = 1$ in the sample, and let $n_1(\bar{s})$ denote the number of nodes with value 1 among the nodes not in the sample. Note that

$n_1(s)$ is observed and $n_1(\bar{s})$ is an unobserved quantity to be estimated or predicted. The realized proportion of value -1 nodes in the population is denoted $Z = (n_1(s) + n_1(\bar{s}))/N$, where N is the total number of nodes in the population.

For a one-wave snowball design with an ignorable initial sample from which all links are traced and with the nondirected stochastic block model, the joint predictive likelihood is

$$\begin{aligned} f(d,\, n_1(\bar{s}) \mid \theta_0,\, \beta_0,\, \beta_1,\, \beta_2) = \\ & p(s \mid y_s,\, a_{s_0 U}) \binom{n(\bar{s})}{n_1(\bar{s})} \\ & \theta_0^{n_0(s)+n_0(\bar{s})} (1-\theta_0)^{n_1(s)+n_1(\bar{s})} \\ & \beta_0^{r_{02}(s_0,\, s)} (1-\beta_0)^{r_{00}(s_0,\, s)+n_0(s_0)n_0(\bar{s})} \\ & \beta_1^{r_{12}(s_0,\, s)} (1-\beta_1)^{r_{10}(s_0,\, s)+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})} \\ & \beta_2^{r_{22}(s_0,\, s)} (1-\beta_2)^{r_{20}(s_0,\, s)+n_1(s_0)n_1(\bar{s})}. \end{aligned} \tag{7}$$

With joint likelihood (7) and independent beta priors and carrying out the integration, the posterior predictive density for the finite-population proportion Z becomes

$$\begin{aligned} f(n_1(\bar{s}) \mid d) \propto & \binom{n(\bar{s})}{n_1(\bar{s})} \\ & B\begin{bmatrix} n_0(s)+n_0(\bar{s})+a,\; n_1(s) \\ +n_1(\bar{s})+b \end{bmatrix} \\ & B[r_{02}+c,\; r_{00}+n_0(s_0)n_0(\bar{s})+d] \\ & B[r_{12}+e,\; r_{10}+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})+f] \\ & B[r_{22}+g,\; r_{20}+n_1(s_0)n_1(\bar{s})+h]. \end{aligned}$$

The Bayes predictor of $n_1(\bar{s})$ is

$$E[n_1(\bar{s}) \mid d] = \sum_{n_1(\bar{s})=0}^{n(\bar{s})} n_1(\bar{s})\, f(n_1(\bar{s}) \mid d).$$

Since $i\binom{n}{i} = n\binom{n-1}{i-1}$,

$$\begin{aligned} E[n_1(\bar{s}) \mid d] \propto & n(\bar{s}) \sum_{i=1}^{n(\bar{s})} \binom{n(\bar{s})-1}{i-1} \\ & B[n_0(s)+n(\bar{s})-i+a,\; n_1(s)+i+b] \\ & B[r_{02}+c,\; r_{00}+n_0(s_0)(n(\bar{s})-i)+d] \\ & B[r_{12}+e, r_{10}+n_0(s_0)i+n_1(s_0)(n(\bar{s})-i)+f] \\ & B[r_{22}+g, r_{20}+n_1(s_0)i+h] \\ & = M_3. \end{aligned}$$

in which M_3 is defined to be the right hand side. Thus, since $M_1 = f(d)$ defined earlier is the proportionality constant, $E[n_1(\bar{s}) \mid d] = M_3/M_1$.

Therefore, the Bayes predictor $\hat{Z}$ of the realized proportion Z of positive nodes in the population is

$$\begin{aligned} \hat{Z} &= E(Z \mid d) = E[(n_1(s)+n_1(\bar{s}))/N \mid d] \\ &= \frac{n_1(s)+(M_3/M_1)}{N}. \end{aligned} \tag{8}$$

4. An Illustrative Example

Here, we consider an example which concerns estimating the percentages of injection drug users and nonusers among a certain target population. Let θ_0 represents the proportion of non injection drug users in the target population. Then $1-\theta_0$ is the proportion of injection drug users. Suppose that there are 200 people in that population. In the first wave sample, 22 people are sampled randomly without replacement and 5 of those sampled are injection drug users whereas 17 are not. The injection drug users are asked to name their injection partners. Note that links are only possible between users and tracing these links can only add users to the sample. The initial users give 12 referrals, of which 10 are distinct users not in the initial sample. The statistics are:

$$\begin{aligned} & n_1(s_0) = 5,\; n_0(s_0) = 17,\; n_1(s) = 15,\; n_0(s) = 17, \\ & n(\bar{s}) = 168,\; r_{22} = 12,\; r_{20} = 93. \end{aligned}$$

In terms of the notation of section 3, $\beta_0 = \lambda_{0011}$ is the probability of a mutual link between two non injection drug users. $\beta_1 = \lambda_{0011} = \lambda_{0111}$ is the probability of a mutual link between injection drug user and non injection drug user (it is natural that the two different orders of node values have the same probability). $\beta_2 = \lambda_{1111}$ is the probability of a mutual link between two injection drug users. Since non injection drug users will be definition not have injection partners, $\beta_0 = \beta_1 = 0$ for this example.

The Bayes estimates for θ_0 and β_2 corresponding to different non-informative priors are given in table 1.

Note that the three non-informative priors are very different from each other. For example, the improper non-informative prior corresponding to $a = b = g = h = 0$ place a lot of its weight on both 0 and 1. This would arise in practice when people in a certain neighborhood are either all injection drug users or are all non injection drug users, but we just do not know which one. On the other hand, the prior corresponding to $a = b = g = h = 1$ place a flat weight to values between 0 and 1. Even though the three priors are very different, the posterior distributions corresponding to these three non-informative priors nearly coincide with each other. Figure 1 shows the posterior distribution of θ_0 and β_2 corresponding to the three non-informative priors. One can conclude that the Bayes estimates here are not sensitive to the specification of the three priors.

Table 1
Bayes Estimates for Noninformative Priors Corresponding to the Specified Values of *a*, *b*, *g*, *h*
(The Values in the Brackets are the 95% HPD Regions)

Bayes estimate	$a = b = g = h = 0$	$a = b = g = h = 0.5$	$a = b = g = h = 1$
$\hat{\theta}_0$	0.7273 (0.5706, 0.8713)	0.7285 (0.5747, 0.8670)	0.7295 (0.5786, 0.8686)
$\hat{\beta}_2$	0.0420 (0.0153, 0.0738)	0.0439 (0.0164, 0.0766)	0.0458 (0.0175, 0.0791)

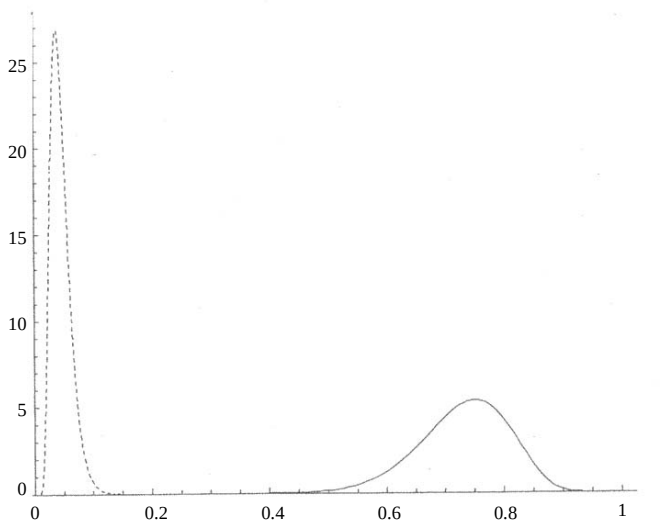


Figure 1. Marginal Posterior distributions: solid line for θ_0 and dashed line for β_2. (The posterior distributions corresponding to the three non-informative priors are given here and they nearly coincide).

For comparison purposes, it is of interest to note that the maximum likelihood estimates obtained using the likelihood function given in (3) are calculated to be: $\hat{\theta}_0 = 0.7604$, $\hat{\beta}_2 = 0.0501$, not far from the Bayes estimates. However, it is not easy to compute confidence intervals for the maximum likelihood estimate whereas one can obtain the posterior intervals for the Bayes estimates without any additional difficulty. For example, a $(1-\alpha)$ highest posterior density (HPD) region can be obtained for the specified α value for each parameter θ_0, β_0, β_1, β_2, where HPD is the region of values that contains $(1-\alpha)$ of the posterior probability for that parameter with the characteristic that the density within the region is never lower than that outside. It is worthwhile to note that the posterior intervals can be directly regarded as having the stated probability of containing the unknown quantity in contrast to the repeated sampling property of frequentist confidence interval. See Gelman, Carlin, Stern and Rubin (1995, pages 104–106) for a discussion on the frequency property of some Bayesian procedures.

From Table 1, we can see that even though the width of the HPD interval of β_2 is large compared to the magnitude of its Bayes estimate, it gives us a rough order-of-magnitude estimate of β_2 and provides useful information to the subject matter specialists.

5. An Empirical Example and Discussion

To examine the properties of estimators and predictors under repeated sampling, socially-networked data from the Colorado Springs study on the heterosexual transmission of HIV/AIDS was used as an empirical population from which to repeatedly sample. The Colorado Springs study, which is described in Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth and Reynolds (1993); Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl (1995), and Darrow, Potterat, Rothenberg, Woodhouse, Muth and Klovdahl (1999), involved a very thorough investigation of a population of people thought to be at high risk for infection with the human immunodeficiency virus. In the study, data were obtained not only on the risk-related behaviors of individuals, but also on their social relationships with other individuals. Risk-related behaviors included various sexual and drug-use behaviors, and the social links examined included sexual and drug-use relationships. Over the course of the study, data were obtained on several thousand people.

For our empirical population we have used the 595 individuals in the study for which the data on both individual risk-related behaviors and relationships to other people in the study are complete. For the node variable of interest we chose a high-risk sexual behavior (commercial sex work) and sexual relationship for the link variable of interest. Figure 2 shows a graphical representation of the empirical population, in which the nodes or circles represent people in the study and the lines represent sexual relationships between pairs of individuals. Presence of the high-risk sexual behavior $(y = 1)$ is indicated by a dark colored circle, while presence of a sexual relationship between two individuals is indicated by a line between the two circles. The positioning of the nodes in the graph is arbitrary, but has been arranged to separate connected components. The largest connected component contains 219 of the 595 people in the population. The next largest connected component contains 12 people, followed by several components of 4, 3 and 2 people. There are 267 people without sexual relationships to others among the 595 in the empirical population. The extremely uneven distribution of connected component sizes exemplified by this population presents one of the challenges to sampling design and inference in such populations.

population

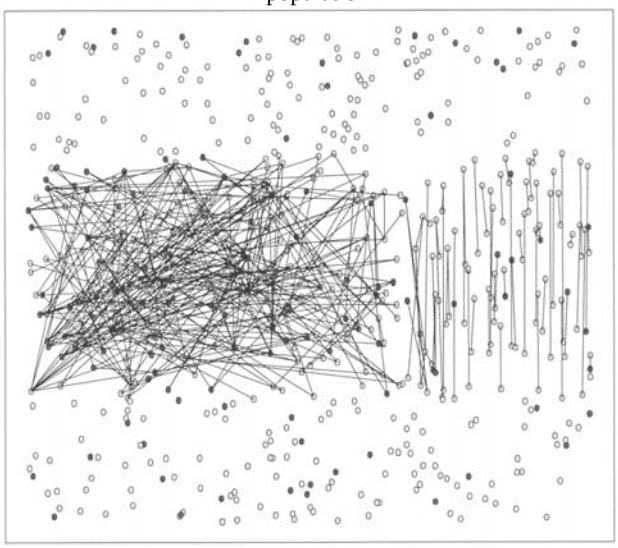


Figure 2. Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat *et al.* 1991; Rothenberg *et al.* 1993; Darrow *et al.* 1999): The 595 people in the empirical population. Dark circles represent individuals with high-risk sexual behavior (sex work). Links between circles indicate sexual relationships.

one-wave snowball sample

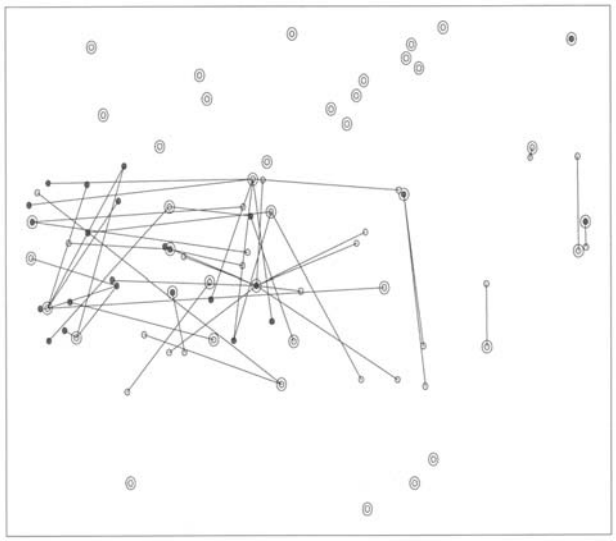


Figure 3. A one-wave snowball sample selected from the Colorado Springs empirical population. From an initial random sample of 40 individuals (circled), links are traced to add one wave of new individuals to the sample.

Figure 3 shows a one-wave snowball sample from this population. First, a simple random sample of 40 nodes (circled in the figure) is selected. All links from these initial nodes are traced to add the additional nodes to the sample.

Repeated sampling of the empirical population was carried out using the one-wave snowball design with initial simple random sample of 40 individuals. The addition of a wave of new nodes brought the total sample size to 85, on average. For each sample, various estimators of the proportion of high-risk individuals $(y=1)$ in the population were computed, and this procedure was repeated 1,000 times. The undirected stochastic block graph model was used for the maximum likelihood and Bayes estimators of θ and the Bayes predictor of the finite-population proportion z. A uniform prior was used for the Bayes procedures. Table 2 and Figure 4 summarize the properties under the repeated sampling of the different estimators. The actual proportion of nodes having value $(y=1)$ in the empirical population is 0.2235. The sample proportion overestimates relative to the actual proportion because the linktracing has a tendency to enrich the sample with high-risk nodes. Each of the model-based estimators has relatively little bias with the link-tracing design.

Table 2
Means and mean square errors of estimators of the population mean of the node values, for the Colorado Springs empirical population. The actual mean of node values in the population is 0.2235294. The design is a one-wave snowball sample with an initial random sample of 40 nodes. The average final sample size was 82.65. The number of simulation runs is 1,000

Type of estimator:	sample proportion	m.l.e.	Bayes estimator	Bayes predictor
mean:	0.3147	0.2155	0.251	0.2142
m.s.e.:	0.011391	0.003279	0.003261	0.003275

In this paper, we employ a Bayesian approach to the estimation problem with link-tracing design and show that, corresponding to the independent beta priors, the posterior distribution can be evaluated analytically. If a more general prior is desired then one can use the Markov Chain Monte Carlo (MCMC) method to evaluate the posterior for that general prior. References for using MCMC techniques in Bayesian computations include Gilks, Richardson and Spiegelhalter (1996) and Gelman, Carlin, Stern and Rubin (1995). The approach used in Gelfand and Smith (1990) can be adapted for the implementation of the MCMC simulations here.

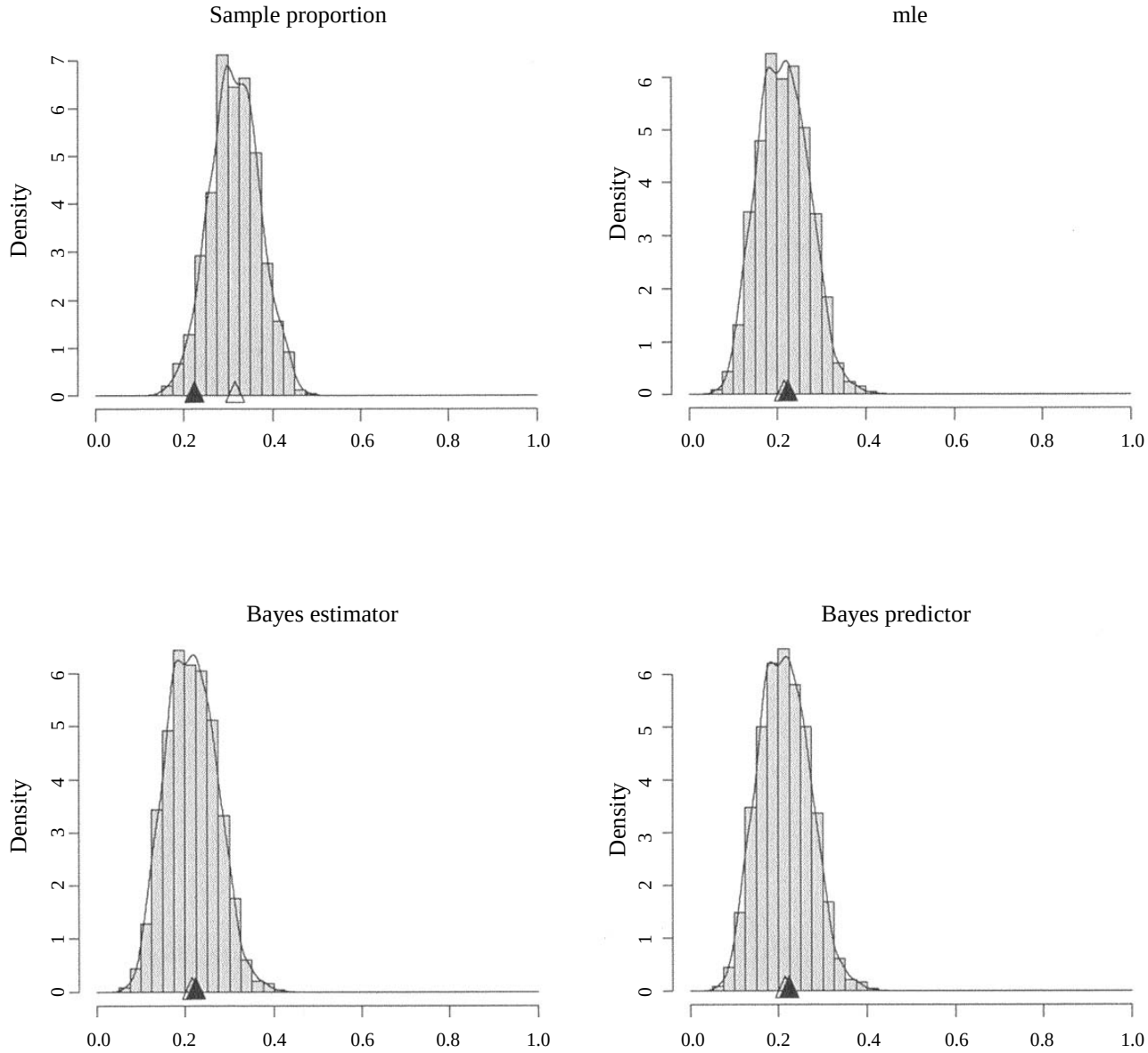


Figure 4. Distributions of estimators of the proportion of individuals in the high-risk category in the Colorado Springs empirical population, with the one-wave snowball design using an initial sample of 40. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. The number of simulations was 1,000.

Acknowledgements

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102), and the National Institutes of Health (R01-DA09872). The authors would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study. We would also like to thank the Associated Editor and the referees for their insightful comments and suggestions.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.) Berlin: Springer-Verlag.

Darrow, W.W., Potterat, J.J., Rothenberg, R.B., Woodhouse, D.E., Muth, S.Q. and Klovdahl, A.S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus*, 32, 143-158.

Erickson, B. (1978). Some problems of inference from chain data. In *Sociological Methodology*, 1979, K.F. Schuessler (Ed.) San Fancisco: Jossey-Bass. 276-302.

Frank, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.

Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-246.

Frank, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, 4, 178-180.

Frank, O. (1977c). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

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

Frank, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*. (P.W. Holland, and S. Leinhardt, Eds.). New York: Academic Press. 319-348.

Frank, O. (1980). Sampling and inference in a population graph. *International Statistical Review*, 48, 33-41.

Frank, O. (1997). Composition and structure of social networks. *Mathematiques, Informatique et Sciences humaines*, 35, 11-23.

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

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Gilk, W.R., Richardson, S. and Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 20, 572-579.

Haldane, J.B.S. (1931). A note on inverse probability. *Proc. Cambridge Philos. Soc.*, 28, 55-61.

Neaigus, A., Frideman, S.R., Goldstein, M.F., Ildefonso, G., Curtis, R. and Jose, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In *Social Networks, Drug Abuse, and HIV Transmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 20-37.

Neaigus, A., Friedman, S.R., Jose, B., Goldstein, M.F., Curtis, R., Ildefonso, G. and Des Jarlais, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 499-509.

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.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks, Drug Abuse, and HIV Tansmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 3-19.

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

Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight. *Bulletin de Methodologie Sociogique*, 36, 59-70.

Snijders, T.A.B., and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14, 75-100.

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: John Wiley & Sons, Inc.