A BAYESIAN APPROACH TO ESTIMATION WITH LINK-TRACING SAMPLING DESIGNS

Mosuk Chow, Steven K. Thompson, The Pennsylvania State University
Mosuk Chow, Department of Statistics,
The Pennsylvania State University, 326 Thomas Building,
University Park, PA 16802, USA.

Key Words: Snowball Samples, Adaptive Sampling, Graph Sampling, Link-Tracing Designs, Network Sampling, Bayesian estimates, Beta prior, Markov Chain Monte Carlo method.

Abstract:
For inference from link-tracing designs, Frank and Thompson (1998) derived the likelihood function for the graph model. In addition, they provided the likelihood functions considered under the symmetric model and also an asymmetric model. In that paper, they used maximum likelihood estimators to estimate the graph model parameters. Here, we propose a Bayesian approach for the estimation problem. For problems with sampling designs that follow social links from one person to another, it is quite often that prior information is available on the proportions that one wants to estimate. Thus, using these information effectively via a Bayesian approach should yield better estimators. Also, under the Bayesian setup, obtaining interval estimates and assessing the accuracy (posterior variances) of the estimators can be done without much added difficulties whereas such tasks would be very 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 large number of questions can be answered simultaneously.

1 Introduction

Social network data include measurements on the relationships between social entities. 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 considered the most practical way to obtain a sample large enough to study. For example, in a study of injection drug use in relation to the spread of the HIV infection, initial respondents may be asked to identify drug-injection partners who are then added to the sample.

Social entities with social structure are often modeled as graphs, with the nodes of the graph representing social entities and the arcs of the graph representing social links, relationships, or transactions. 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, 1979a, 1979b, 1980, 1997) has many important results in sampling for social networks. His classic work (Frank 1971) presents a basic solution for estimating graph quantities from the sample data. Snijders and Nowicki (1997) proposes various statistical approaches, including a Bayesian approach, to estimation and prediction for stochastic blockmodels for graphs with latent block structure.

Snowball sampling is one type of link-tracing sampling design in which individuals in an initial sample were asked to identify a fixed number of acquaintances, who in turn were asked to identify the same number of acquaintances and so on for a fixed number of stages or waves. This very clever network sampling idea originated from Goodman (1961). Erickson (1978) and Frank (1979b) review snowball sampling design 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 1990, 1997, 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. Design-based estimation methods have the advantage that properties such as design-unbiasedness or consistency do not depend for their validity on any assumed model for the population. On the other hand, these properties do depend on the sampling design being carried out as specified. In this paper, we use the model-based methods described in Thompson and Frank (1998).

The model-based methods do depend on the assumed model for the population or graph. Their advantage is that they apply to a wide range of sample selection procedures. In many real 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. For example, in a long-term study on the heterosexual transmission of HIV infection (Rothenberg, et al., 1995), the target population of interest consisted of commercial sex workers, their paying and nonpaying partners, persons who use injectable drugs, and the sexual partners of drug users in the Colorado Springs area. Persons in the purposively-selected initial sample were interviewed and, in addition to their individual characteristics, identities of their sexual partners were obtained. Persons named by two or more respondents were also located and interviewed. The wide range of link-tracing procedures used in studies such as these has motivated the emphasis in this paper on model-based inference methods.

In Thompson and Frank (1998), maximum likelihood estimators of population graph parameters and predictors of realized population graph quantities were described and compared to commonly-used conventional estimates or data summaries such as sample means and proportions of node or link values. Design-based estimation methods have the advantage that properties such as design-unbiasedness or consistency do not depend for their validity on any assumed model for the population. On the other hand, these properties do depend on the sampling design being carried out as specified. In this paper, we use the model-based methods described in Thompson and Frank (1998). The model-based methods do depend on the assumed model for the population or graph. Their advantage is that they apply to a wide range of sample selection procedures. In many real 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. For example, in a long-term study on the heterosexual transmission of HIV infection (Rothenberg, et al., 1995), the target population of interest consisted of commercial sex workers, their paying and nonpaying partners, persons who use injectable drugs, and the sexual partners of drug users in the Colorado Springs area. Persons in the purposively-selected initial sample were interviewed and, in addition to their individual characteristics, identities of their sexual partners were obtained. Persons named by two or more respondents were also located and interviewed. The wide range of link-tracing procedures used in studies such as these has motivated the emphasis in this paper on model-based inference methods.

In Thompson and Frank (1998), maximum likelihood estimators of population graph parameters and predictors of realized population graph quantities were described and compared to commonly-used conventional estimates or data summaries such as sample means and proportions of node or link values. It is important to note that in most cases the conventional estimates are not the best estimates and conventional estimates or unadjusted summaries of sample data obtained through link-tracing procedures can be misleading if viewed as pertaining to population or whole graph characteristics.

In this paper, we consider a Bayesian approach for the estimation problem. For real problems with sampling designs that follow social links from one person to another, it is quite often the case that there is prior information on the characteristics that one wants to estimate. Thus, using this information effectively via a Bayesian approach should yield better estimators. Also, note that under the Bayesian setup, obtaining interval estimates and assessing the accuracy (posterior variances) of the estimators can be done without much added difficulties whereas such tasks would be very 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 large number of questions can be answered simultaneously. The prediction problem, which is an important problem in studies of hidden and hard-to-access human populations, can also be answered once we obtain the posterior distribution.

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 sample obtained from a link-tracing designs will be presented and a Bayesian inference method will be introduced.

2 Notations and likelihood for a Full Graph Model with Links Related to Node Values

Consider a graph of $N$ nodes labeled $1, 2, ..., N$. Associated with the $u$th node is a variable of interest $Y_u$. The full set of node labels is denoted by $U = \{1, 2, ..., N\}$ and the sequence of node variables by $Y = (Y_1, ..., Y_N)$. For two distinct nodes $u$ and $v$, the indicator variable $X_{uv}$ equals one if there is an arc (directional link) from $u$ to $v$ and zero otherwise. The matrix of arc indicators, having $X_{uv}$ as the element in the $u$th row and $v$th column, is the graph adjacency matrix, denoted $X$. For convenience we will assume that the diagonal elements $X_{uu}$ are zero. The ordered pair $(u, v)$ is referred to as a dyad of type $(Y_u, Y_v, X_{uv}, X_{vu})$. In the following assumed model the node variables $Y_1, ..., Y_N$ are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_i = 1) = \theta_i$, for $i = 0, 1$, and $\theta_0 + \theta_1 = 1$. Conditional on the node values $Y_1, ..., Y_N$, the dyads $(X_{uv}, X_{vu})$ are independent, for $1 \leq u < v \leq N$, with conditional distribution given by $P((X_{uv}, X_{vu}) = (k, l) | Y_u = i, Y_v = j) = \lambda_{ijkl}$ 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 $\sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order
to get graph probabilities not depending on node identities, the following symmetry requirements are needed: \( A_{1100} = A_{1011}, A_{1010} = A_{0110}, A_{1101} = A_{0101}, A_{1001} = A_{0110}, A_{0010} = A_{0001} \) and \( A_{1000} = A_{0100} \). 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 \((Y_u, Y_v; X_u, X_v) = (ijkl)\). The likelihood for the full graph under the model with parameters \((\theta, \lambda)\) is

\[
L(\theta, \lambda; Y, X) = \prod_{i=0}^{1} \theta_i^{N_i} \prod_{i=0}^{1} \prod_{j=0}^{1} \prod_{k=0}^{1} \prod_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}}
\]

(1)

3 Bayesian Inference from Link-Tracing Designs

3.1 Likelihood Functions for the Sample

A sample \( s \) from the graph is a subset of nodes and a subset of node pairs. The sample data \( d = (s, y_s, x_s) \) are a function of the sample selected and of the graph values \( y \) and \( x \). For any design in which the selection of the sample depends on graph \( y \) and \( x \) values only through those values \( y_s \) and \( x_s \) 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.

Consider a link-tracing design without subsampling of links. An initial sample \( s_0 \) is selected and links out from nodes in \( s_0 \) are followed to add the set \( s_1 \) of nodes not in \( s_0 \) that are adjacent after nodes in \( s_0 \). The whole sample is \( s = s_0 \cup s_1 \). The entire set of labels 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 only 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 \( P(s|y, x_{so}) \), since the selection procedure depends on both node and link values. The data are \( d = (s, y_s, x_{so}) \) so that the design depends on \( y \) and \( x \) values only through those in the data and is thus ignorable.

With the graph model described in the previous section, it then follows from Thompson and Frank(1998) that the likelihood with the sample data is given by:

\[
L(\theta, \lambda; d) = P(s|y, x_{so}) \prod_{u=1}^{N} \theta_{y_u} (\prod_{u<v}^{N} \lambda_{y_u y_v x_{uv}})
\]

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

For the link-tracing designs in which all links, rather than a subsample from the initial samples are traced, all of the elements in the submatrix \( x_{so} \) are zero. It has been shown by Thompson and Frank (1998) that the likelihood function can then be written as:

\[
L(\theta, \lambda; d) = P(s|y, x_{so})(\prod_{i=0}^{1} N_i^{n_i(s)})
\]

\[
\times \prod_{ij} \lambda_{ij}^{m_{ijkl}(s_0, s_1)}(\prod_{ijk} \lambda_{ijkl}^{m_{ijkl}(s_0, s_1)})
\]

\[
\times \prod_{i} \theta_i^{n_i(s)} x_{so}^{\lambda_i(s)}
\]

(2)

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. Also, \( m_{ijkl}(s_0, s_0) \), \( m_{ijkl}(s_0, s_1) \) 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 \( \lambda_{ijkk} = \lambda_{jikk} \) for \( i, j, k = 0, 1 \), with \( \lambda_{ij00} + \lambda_{ij11} = 1 \). Let \( \beta_{ij+} = \lambda_{ij11} \),

\[
r_{i+j,k+l} = m_{ijkl}(s_0, s), \quad r_{i+0} = m_{1000}(s_0, s) + m_{1001}(s_0, s), \quad r_{12} = m_{0111}(s_0, s) + m_{1011}(s_0, s),
\]

the above expression can be rewritten as:

\[
L(\theta, \beta; d) = P(s|y, x_{so}) \prod_{o}^{n_o(s)} (1 - \theta_o)^{n_{o}(s)} \beta_{o}^{\alpha_o, 2}
\]

\[
\times (1 - \beta_o)^{\alpha_o} \beta_o^{\alpha_o, 2} (1 - \beta_0)^{\gamma_{0,1}} \beta_o^{\gamma_{0,1}, 2}
\]

\[
\times (1 - \beta_2)^{\alpha_2} \theta_2 (1 - \beta_o)^{n_{o}(s_o)} (1 - \beta_1)^{n_{1}(s_o)}
\]

\[
+ \theta_1 (1 - \beta_1) n_{1}(s_o) (1 - \beta_2)^{n_{2}(s_o)}
\]

(3)

613
3.2 Choice of Prior Distributions

Since there are no specific constraints on $\theta_o, \beta_o, \beta_1, \beta_2$, we may assume independent priors on $\theta_o, \beta_o, \beta_1, \beta_2$, 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 distribution on $[0, 1]$ can be well approximated by some beta distribution and the class of beta distribution is reasonably rich to model the uncertainty about the parameter. Also, the expression in (5) is quite complex but beta priors can yield a tractable posterior distribution. Using beta priors, we obtain an analytic formula for the Bayes estimates and a computer program in Mathematica has been written to evaluate them. In addition, we wrote a FORTRAN program, employing the Markov Chain Monte Carlo method to evaluate the Bayes estimates. The MCMC method is a simulation technique which can be used to compute Bayes estimates for general prior distributions specified by an user. From the MCMC method, we obtain a random sample from the posterior distribution which is used to answer a large number of inference questions simultaneously whereas the analytic formula can only provide point estimates. However, one can use results from the Mathematica program to check the accuracy of the MCMC method under the beta priors since it may take a while for the MCMC results to converge. The prior considered in this paper consists of independent beta priors for the parameters:

\[
\begin{align*}
\pi(\theta_o, \beta_o, \beta_1, \beta_2) & \propto \theta_o^{a-1}(1 - \theta_o)^{b-1}\beta_o^{c-1} \\
& \quad (1 - \beta_o)^{d-1}\beta_1^{e-1}(1 - \beta_1)^{f-1} \\
& \quad \beta_2^{g-1}(1 - \beta_2)^{h-1}
\end{align*}
\]

In determining the constants $a$ and $b$ it is often useful to equate the mean

\[
E[\theta_o] = a/(a + b)
\]

of Beta($a, b$) to a value which represents your belief about the location of $\theta_o$ and the variance

\[
Var[\theta_o] = \frac{ab}{(a + b)^2(a + b + 1)}
\]

of Beta($a, b$) to a value which represents the uncertainty you put on your specified $\theta_o$ 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 they share needle with. If the value $y_u = 1$ represents a user of injection drugs, then $\theta_o$ is the percentage of non-injection drug users in that community. Quite often an investigator may be able to provide an estimate of its central location and the corresponding spread.

In the case of complete ignorance, noninformative priors will be appropriate. There are three commonly used noninformative priors. The first one is the uniform prior, which corresponds to Beta(1,1). The second one, Beta(0,0) has an improper density. It is equivalent (by the usual change of variable argument) to a prior uniform in the log-odds log($\theta_o/(1 - \theta_o)$). A possible compromise between Beta(1,1) and Beta(0,0) is Beta(1/2, 1/2), which has a proper density. This prior implies a uniform prior for $\sin^{-1}\sqrt{\theta_o}$. In this paper, we are going to consider all of these three priors and provide a comparison of the resulting Bayes estimates. An excellent discussion on the noninformative priors is given in Berger(1985, p.89-90).

3.3 Posterior Distribution and Bayes Estimates

In the Bayesian framework, one takes prior beliefs about the unknown parameters and then modifies these prior beliefs in the light of relevant data which one has collected to arrive at posterior beliefs. The posterior distribution of $\theta_o, \beta_o, \beta_1, \beta_2$ given data $d$ will be denoted $\pi(\theta_o, \beta_o, \beta_1, \beta_2|d)$. It combines the prior beliefs about the unknown parameters with the information about the parameters contained in the data to give a composite picture of the final beliefs about them. In our problem, the posterior distribution corresponding to the beta priors is given by:

\[
\begin{align*}
\pi(\theta_o, \beta_o, \beta_1, \beta_2|d) & \propto P(s|y, \tau_{s_o})^\theta_o^{n_o(s_o)+a-1}(1 - \theta_o)^{n_1(s_o)+b-1} \\
& \quad \beta_o^{c+o-1}(1 - \beta_o)^{d+1}\beta_1^{e+1}(1 - \beta_1)^{f+1} \\
& \quad \beta_2^{g+1}(1 - \beta_2)^{h+1}
\end{align*}
\]

Let

\[
\begin{align*}
\theta_o(1 - \beta_o)^{n_o(s_o)}(1 - \beta_1)^{n_1(s_o)} + \\
\theta_1(1 - \beta_1)^{n_o(s_o)}(1 - \beta_2)^{n_1(s_o)}]
\end{align*}
\]

614
3.4 Markov Chain Monte Carlo method

The Markov chain Monte Carlo (MCMC) method is essentially Monte Carlo integration using Markov chains. Monte Carlo integration draws samples from the required distribution, and then forms sample averages to approximate expectations. MCMC draws these samples by running a cleverly constructed Markov Chain for a long time. There are many ways of constructing these chains, but all of them, including the Gibbs sampler (Geman and Geman, 1984), are special cases of the general framework of Metropolis et al. (1953) and Hastings (1970). The work of Geman and Geman (1984) led to the introduction of MCMC into mainstream statistics via the articles by Gelfand and Smith (1990) and Gelfand et al. (1990). The book by Gilks, Richardson and Spiegelhalter (1995) is an excellent reference for using MCMC techniques in Bayesian computations.

For the problem of computing the posterior distribution of $\pi(\theta, \beta_0, \beta_1, \beta_2 | d)$ given in (6), we perform a Monte Carlo simulation analysis, combining aspects of Gibbs sampling (Gelfand and Smith 1990) and sampling-importance-resampling (Rubin 1988) to evaluate the posterior estimates.

References


