



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium
Series - Proceedings**

**Symposium 2004: Innovative
Methods for Surveying
Difficult-to-reach Populations**

2004



NETWORK SAMPLING WITH A BAYESIAN APPROACH

Mosuk Chow and Steven K. Thompson¹

ABSTRACT

In surveying difficult-to-reach populations, network sampling is sometimes the only practical way to obtain a sample large enough for an effective study. In this paper, we consider the problem of estimating social network properties from samples. A link-tracing sampling design is considered. This paper describes the framework of using a Bayesian approach for the estimation of social network properties when the population with its social network structure is modeled as a stochastic graph. A Mathematica program is provided in this paper to compute the Bayes estimates of an illustrative example.

KEYWORDS: Bayesian Estimation; Beta Priors; Link-Tracing Designs; Model-Based Approach; Network Sampling.

1. INTRODUCTION

1.1 Background

Links between people or other social entities are commonly studied in terms of networks. Social network data include measurements on the relationships between people or other social entities as well as measurements on entities themselves. In link-tracing designs, social links are followed from one respondent to another to obtain the sample. The method of network sampling was first used because it was unavoidable when unitary counting rules were hard to define since the populations of interest were linked to a multitude of establishments. In 1960's, there was a big medical provider survey of estimating the prevalence of a life threatening disease, cystic fibrosis, in three New England states (Kramm, et al 1962). Cystic fibrosis is a very severe disease and the patients usually went to more than one medical center to get treatment. Birnbaum and Sirken at National Center of Health Statistics (NCHS) proposed several unbiased estimators, including the later widely used multiplicity estimator and Horvitz-Thompson network estimators, to deal with the problem in Birnbaum and Sirken (1965). After that, the potential of using network sampling to increase the "yield" of household surveys of rare characteristics and to obtain estimates with smaller sampling errors than conventional procedures is recognized. For example, Nathan (1976) utilized the multiplicity rules to estimate the number of births in Israel. For hidden and hard-to-access human populations, such sampling designs are sometimes the only practical way to obtain a sample large enough for an effective study. One such example is study of risky human behavior. Thompson and Collins (2001) formulates adaptive sampling in the graph setting as well as the spatial setting in research on risk-related behaviors.

There are many forms of link-tracing designs. One frequently used one is snowball sampling, first termed in Goodman (1961). Snowball sampling is one type of link-tracing sampling design in which individuals in an initial sample were asked to identify acquaintances, who in turn were asked to identify acquaintances, and so on for a fixed number of stages or waves. It is important to be able to estimate network properties from samples. The term "snowball sampling" was used in Snijders (1992) to include designs in which only a subsample of links from each node is traced. Frank and Snijder (1994) consider model and design based estimation of a hidden population size based on snowball sampling. Thompson and Frank (2000) use a model-based approach to inference with link-

¹ Mosuk Chow, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A.; Steven K. Thompson, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A. and Los Alamos National Laboratory, Los Alamos, NM 87544, U.S.A.

tracing designs. In their paper, they derive the likelihood functions considered under the symmetric model and also an asymmetric model. They then use maximum likelihood estimators to estimate the graph model parameters and predictors of realized population graph quantities. The maximum likelihood estimator was compared to conventional estimates and data summary statistics.

1.2 Bayesian approach to network analysis

Recently, there has been plenty of interest in using Bayesian approach to network analysis. We mention a few of the Bayesian papers here. Wong (1987) considers the p_I model proposed by Holland and Leinhardt (1981) for analyzing digraphs that arise in studies of networks. In the paper, a Bayesian approach, using an exchangeable normal prior on the parameters representing the attractiveness and expansiveness characteristics of the nodes, is proposed. Snijders and Nowicki (1997) propose various statistical approaches, including a Bayesian approach, for estimation and prediction with stochastic blockmodels for graphs with latent block structure. More recently, Nowicki and Snijders (2001) propose a statistical approach to a posteriori blockmodeling for digraphs. The probability model assumes that the vertices of the digraph are partitioned into several unobserved classes and that the probability of the relation between two vertices depends only on the classes to which they belong. A Bayesian estimator based on Gibbs sampling is proposed in the paper. In Hoff, Raftery and Handcock (2002), a class of models where the probability of a relation between actors depends on the positions of individuals in an unobserved “social space” is developed. They make inference for the social space within maximum likelihood and Bayesian frameworks, and propose Markov Chain Monte Carlo procedures for making inference on latent positions and the effects of observed covariates.

1.3 Bayesian approach to Link-tracing designs

As we have mentioned in section 1.1, Thompson and Frank (2000) derive the likelihood function for the graph model under the symmetric model and also an asymmetric model. In Chow and Thompson (2003), a Bayesian approach for the estimation and prediction problems when the population with its social network structure is modeled as a stochastic graph was proposed. In this paper, we provide a Mathematica program to evaluate the illustrative example provided in that paper. We want to show that the methodology is accessible to the practitioner and the computation is simple and fast. The Bayesian approach can be adopted widely to deal with these problems.

2. NOTATIONS

2.1 Likelihood function of a full graph model

For a graph of N nodes labelled $1, 2, \dots, N$, and the variable of interest Y_u associated with the u th node, we denote the full set of node labels by $U = \{1, 2, \dots, N\}$ and $Y = (Y_1, Y_2, \dots, Y_N)$. The indicator variable A_{uv} is equal to 1 if there is a directional link from u to v and to zero otherwise where u, v are two distinct nodes. As in Chow and Thompson (2003), we assume that Y_1, Y_2, \dots, Y_N are independent, identically distributed (i.i.d) Bernoulli random variables and $P(Y_u = 0) = \theta_0$ and $P(Y_u = 1) = \theta_1$ with $\theta_0 + \theta_1 = 1$. Conditional on the node values (Y_1, Y_2, \dots, Y_N) , the dyads (A_{uv}, A_{vu}) are independent, for $1 \leq u < v \leq N$. The conditional probability is given by $P[(A_{uv}, A_{vu}) = (k, l) | Y_u = i, Y_v = j] = \lambda_{ijkl}$ 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 $\lambda_{ij\cdot\cdot}$ and equals to 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 $\lambda_{1000} = \lambda_{0100}$. Let N_i denote the total number of nodes with value i in the graph so that $N_0 + N_1 = N$. Denote the total number of dyads of type (i, j, k, l) by M_{ijkl} . Then the likelihood for the full graph under the model with parameters (θ, λ) is

$$L(\theta, \lambda; Y, A) = \left(\prod_i \theta_i^{N_i} \right) \left(\prod_i \prod_j \prod_k \prod_l \lambda_{ijkl}^{M_{ijkl}} \right).$$

2.2 The likelihood function given the sample data of the full symmetric model

The Bayesian methodology is demonstrated for the full symmetric model. The full symmetric model has $\lambda_{ijkl} = 0$ for $k \neq l$, $\lambda_{ijkk} = \lambda_{ijkk}$ for $i, j, k = 0, 1$ and $\lambda_{ij00} + \lambda_{ij11} = 1$. Denote by $\beta_{i+j} = \lambda_{ij11}$ and thus β_k denotes the probability of a mutual link between two nodes having total value k where k may take values 0, 1 or 2. Consider the link tracing design in which an initial sample s_0 is selected and all links 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. In this paper and in Chow and Thompson (2003), the design in which the decision to follow the links from node u depends on the node value y_u and the design can be written as $P(s|y_s, a_{s_0U})$. The data are $d = (s, y_s, a_{s_0U})$. 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 the Bayes inference depends only on the assumed model. Let $n_i(s)$, $n_i(s_0)$ 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} . Let r_{ij} denote the 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 likelihood function is given by:

$$L(\theta, \beta; d) = P(s|y_s, a_{s_0U}) \theta_0^{n_0(s)} (1 - \theta_0)^{n_1(s)} \beta_0^{r_{0,2}} (1 - \beta_0)^{r_{0,0}} \beta_1^{r_{1,2}} (1 - \beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1 - \beta_2)^{r_{2,0}} [\theta_0 (1 - \beta_0)^{n_0(s_0)} (1 - \beta_1)^{n_1(s_0)} + (1 - \theta_0) (1 - \beta_1)^{n_0(s_0)} (1 - \beta_2)^{n_1(s_0)}]^{n(\bar{s})}$$

3. BAYES ESTIMATORS

3.1 Priors for the parameters

Since the parameters $\theta_0, \beta_0, \beta_1, \beta_2$ take values between 0 and 1, it is quite common to put a beta prior on them. In addition, since there are no specific constraints on these parameters, we may assume independent priors. Note that beta priors is a reasonably rich class of distributions which can approximate most unimodal distributions on $[0,1]$ well. The prior distribution is given by:

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

When there are past knowledge of the mean and variance of these parameters, one can solve for the values of a, b, c, d, e, f, g, h by equating the corresponding values to the mean of a *Beta* (a, b) = $a/(a+b)$ and the variance of a *Beta* (a, b) = $ab/((a+b)^2(a+b+1))$. In many instances, conducting a pilot study can help to suggest the values of a, b, c, d, e, f, g, h . In the case of no information, one can consider using noninformative priors.

3.2 Posterior Distribution

In Bayesian analysis, in addition to specifying the likelihood function, one also specifies a prior distribution on the unknown parameters. Inference concerning these unknown parameters is then based on the posterior distribution. For our problem and the prior specified above, the posterior distribution is given by:

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

Once we have the posterior distribution, the Bayes estimates for each parameter are the posterior mean and can readily be found. For example, to find the posterior mean of θ_0 , let

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

Since $\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = B(\alpha, \beta)$ is the beta function, we can obtain that:

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

and also that:

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

The Bayes estimate for θ_0 is just:

$$E(\theta_0 | d) = \frac{M_2}{M_1}$$

3.3 A Mathematica program to evaluate the Bayes Estimate

Even though the beta priors are not the conjugate priors for the likelihood function, they do produce posterior means that are relatively easy to evaluate. We have provided an illustrative example together with a Mathematica program here to demonstrate that it is easy to compute the Bayes estimates. We are going to consider estimation of percentage of injection drug users in a certain community. For better presentation, we will let θ_0 denote the proportion of non-injection drug users in that community whereas $1 - \theta_0$ denotes the proportion of injection drug users. To correlate the presentation here with the Chow and Thompson (2003) paper, we will use the same numerical values. If there are 200 people in the community, 22 people are sampled randomly without replacement and 5 of these people are injection drug users whereas 17 are not. The injection drug users are asked to name their injection partners. Since links are only possible between users and tracing these links can only add users to the

sample. The initial users give a total of 12 referrals and 10 of which are distinct users not in the initial sample. Thus:

$$n_1(s_0) = 5, \quad n_0(s_0) = 17, \quad n_1(s) = 15, \quad n_0(s) = 17, \quad n(\bar{s}) = 168,$$

$$r_{22} = 12, \quad r_{20} = 93$$

Note that β_0 is the probability of a mutual link between two and β_1 is the probability of a mutual link between injection drug user and non injection drug. Both these probabilities should be 0. So the two unknown parameters that remain are only θ_0 and β_2 . In this example, β_2 is the probability of a mutual link between two injection drug users. In the Mathematica program we show below, we will compute the Bayes estimates for θ_0 and β_2 corresponding to the noninformative prior $a = b = g = h = 1/2$. It can also be shown that the Bayes estimates corresponding to the other two noninformative priors: $a = b = g = h = 0$ and $a = b = g = h = 1$ have similar values to the following computed ones. One can conclude that for this set of data, the Bayes estimates are not sensitive to the specification of these three priors.

The Mathematica program to evaluate the Bayes estimates of θ_0 and β_2 when the prior is chosen to be the noninformative prior with $a = b = g = h = 1/2$

```

n1s0=5;
n0s0=17;
n1s = 15;
n0s = 17;
nsbar = 168;
r22=12;
r20=93;
a=0.5;
b=0.5;
g=0.5;
h=0.5;
AA= n0s + a;
BB= n1s + b;
GG= r22 + g;
HH= r20 + h;
k = nsbar;
sumzi = Sum[Binomial[k,i]Beta[AA+i,k+BB-i]Beta[GG,HH+(k-i)n1s0],
{i,0,k}];
sumtheta = Sum[Binomial[k,i]Beta[AA+1+i,k+BB-i]Beta[GG,HH+(k-i)n1s0],
{i,0,k}];
thetabayes = sumtheta / sumzi // N
sumbeta = Sum[Binomial[k,i]Beta[AA+i,k+BB-i]Beta[GG+1,HH+(k-i)n1s0],
{i,0,k}];
betabayes = sumbeta / sumzi // N

```

The output will show the Bayes estimate for θ_0 which is called thetabayes in the program. The output will also show the Bayes estimate for β_2 which is called betabayes in the program. For this data set, the result is:

$$thetabayes = .728465, \quad betabayes = .0438741$$

4. CONCLUSION

In this paper, we provide the Mathematica program and show that it is simple and quick to compute the Bayes estimates. For problems using link-tracing designs, it is quite often that there is prior information on the characteristic that one wants to estimate. Incorporating these information is easy in the Bayesian approach and also if one wants to use the program. One just needs to change the values of a , b , g , h in the program to reflect the prior belief. In addition, under the Bayesian setup, obtaining interval estimates and assessing the accuracy of the estimators can be done without much added difficulties.

REFERENCES

- Birnbaum, Z.W. and Sirken, M.G. (1965), "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates", *Vital and Health Statistics*, PHS Publication No1, Series 2, No.11 US Government Printing Office, Washington.
- Chow, M. and Thompson, S. K. (2003), "Estimation with Link-Tracing Sampling Designs – A Bayesian Approach", *Survey Methodology*, 29, no. 2, pp. 197-205.
- Frank, O. and Snijders, T.A.B. (1994), "Estimating the Size of Hidden Populations Using Snowball Sampling", *Journal of Official Statistics*, 10, pp. 53-67.
- Goodman, L.A. (1961), "Snowball Sampling", *Annals of Mathematical Statistics*, 20, pp. 572-579.
- Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002), "Latent Space Approaches to Social Network analysis", *Journal of the American Statistical Association*, 97, pp. 1090-1098.
- Holland, P.W. and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs", *Journal of the American Statistical Association*, 76, pp. 33-65.
- Kramm, E.R., Crance, M.M., Sirken, M.G., and Brown, M.L. (1962), "A Cystic Fibrosis Pilot Survey in Three New England States", *American Journal of Public Health*, 52, pp. 2041-2057.
- Nathan, G. (1976), "An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules", *Journal of the American Statistical Association*, 71, pp. 803-815.
- Nowicki, K. and Snijders, T.A.B. (2001), "Estimation and Prediction for Stochastic Block-structures", *Journal of the American Statistical Association*, 96, pp. 1077-1087.
- Snijders, T.A.B. (1992), "Estimation on the Basis of Snowball Samples: how to weight", *Bulletin de méthodologie sociologique*, 36, pp. 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, pp. 75-100.
- Thompson, S.K. and Collins, L.M. (2001), "Adaptive Sampling in Research on Risk-related Behaviors", *Drug and Alcohol Dependence*, 68, pp. S57-S67.
- Thompson, S. K. and Frank, O. (2000), "Model-based Estimation with Link-Tracing Sampling Designs", *Survey Methodology*, 26, no. 1, pp. 87-98.
- Wong, G.Y. (1987), "Bayesian Models for Directed Graphs", *Journal of the American Statistical Association*, 82, pp. 140-148.