Research Report

Department of Statistics

No. 2003:5

Bayesian Estimation of Blockstructures from Snowball Samples

Christian Tallberg

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Bayesian Estimation of Blockstructures from Snowball Samples

Christian Tallberg

Abstract

The snowball sampling procedure is considered to estimate the size of small hidden populations. Previous work in this area have been based on models where the probability of relations is the same across all pairs of members in the network. Here, we use a more general blockmodel which allows a richer probabilistic structure. Bayesian methods are employed and the posterior distribution of the size of the population is easily computed analytically if the block labels are known. If the block labels are unknown or latent, the posterior distribution is computed by the Gibbs sampler algorithm. The Gibbs sampler also provides us with the posterior distributions of other model parameters without any additional difficulty.

Keywords: Bayesian analysis; Hidden population; Network sampling; Random graphs; Stochastic blockmodels.

1 Introduction

In studies where the purpose is to estimate features in so called hidden human populations, standard probability sampling designs will lead to inefficient estimates if the samples are of moderate size. Often however, contact patterns between members of the population exists, which facilitates for more effective procedures to collect data. Such procedures are link-tracing sampling designs, which means that social links are followed from one respondent to another to obtain a sample.

We shall consider a link-tracing procedure 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*.

The procedure was termed snowball sampling by Goodman (1961). Various statistical methods for snowball samples are investigated by Frank (1979), Snijders (1992), Frank and Snijders (1994), Thompson and Frank (2000) and Spreen and Coumans (2001). Two papers addressing snowball sampling from a Bayesian viewpoint are given by Chow and Thompson (1999) and Tallberg (2003). An overview of link-tracing designs and further references to the literature on the subject is provided in Spreen (1992).

The motivation for this paper is that the models proposed in those articles fail to yield accurate inferences if members of the population have markedly different contact patterns. For example, a non-representative large proportion of members with particularly rich contact structure in the sample can lead to serious overestimation of the population size. The goal of this paper is to propose inference methods that consider different contact patterns among the members, which should yield more effective estimators.

The concept, where members with similar contact patterns and relational features in a population are partitioned into the same groups or blocks, is in the social network literature described as blockmodels. The members and their relational structure can be represented by graphs, where the members are referred to as vertices or actors and the relations are referred to as arcs. By assuming a random graph, we shall consider stochastic blockmodels which implies that members with the same probabilistic relational structures to the other actors in the graph, are partitioned into the same blocks. Papers describing various aspects of blockmodels include White, Boorman and Breiger (1976), Fienberg and Wasserman (1981), Holland, Laskey and Leinhardt (1983) and Wasserman and Anderson (1987) to mention a few. Blockmodels may also be used to model the important concept of centrality by defining the block consisting of members with the largest probability to generate relations as central. Contributions to the centrality concept, are works given by Beauchamp (1965), Höivik and Gleditsch (1975), Freeman (1977, 1979), Nieminen (1974), Snijders (1981) and Frank (2002) among others. For a more basic clarification on blockmodels, centrality and other social network concepts, the reader is referred to Wasserman and Faust (1994).

The Bayesian approach to analyze data has seen an upsurge in popularity in recent years, including the field of social networks. Snijders and Nowicki (1997) and Nowicki and Snijders (2001) use a Bayesian approach to blockmodels, where the probability of a relation between two actors depends only on the blocks to which the actors belong but is independent of the actors. In their setting the population size is known and the aim is to predict block

affiliation, whereas here the size is unknown and the focus of attention.

2 Blockmodels and concepts of snowball sampling

We shall follow the notation outlined in Frank and Snijders (1994). Consider a directed graph on the vertex set $V = \{1, 2, ..., v\}$, and let V^2 denote the set of all ordered pairs (u, w) from V. By denoting the adjacency matrix y, each entry $y_{uw}, (u, w) \in V^2$ takes the value 1 if an arc is present from u to w and 0 otherwise. The diagonal entries of y are equal to 1. We assume that V is partitioned into c mutually exclusive non-empty vertex subsets $V_1, V_2, ..., V_c$ called blocks, where $|V_j| = v_j, j = 1, ..., c$ and $v = v_1 + ... + v_c$. Let z_u be the block label of actor u. Conditional on all $z_1, ..., z_v$, the elements y_{uw} for $u \neq w$ are independent random variables with probability

$$\Pr(y_{uw} = 1 \,|z_1, ..., z_v) = \Pr(y_{uw} = 1 \,|z_u, z_w) = \beta_{z_u, z_w}.$$

Thus, by conditioning on block affiliations, the arcs occur independently, and the probability of an arc from vertex u to vertex w depends on the block labels z_u and z_w only.

The snowball sample $S \subseteq V$ consists of an initial sample $S_0 \subseteq V$ selected by some adequate design and q waves $S_1, ..., S_q \subseteq V$ following the initial sample. Here, we restrict ourselves to only one wave after the initial sample. Since we assume a blockmodel, the initial sample is partitioned into c blocks denoted by $S_{01} = S_0 \cap V_1, S_{02} = S_0 \cap V_2, ..., S_{0c} = S_0 \cap V_c$, where $|S_{0j}| = n_j$ and $n = |S_0| = n_1 + n_2 + ... + n_c$ is the total size of the initial sample. Further, the first wave S_1 is analogously partitioned into c blocks denoted by $S_{11} = S_1 \cap V_1, S_{12} = S_1 \cap V_2, ..., S_{1c} = S_1 \cap V_c$, where $|S_{1j}| = m_j$ and $m = |S_1| = m_1 + m_2 + ... + m_c$ is the total size of the first wave.

The subset of vertices that are adjacent from vertex w is denoted by $A_w = \{u \in V : y_{wu} = 1\}$, and given by row w of y. The size of A_w, $|A_w|$, is called the out-degree and is denoted by a_w. If we by

$$A(S) = \bigcup_{w \in S} A_w$$

denote the subset of vertices adjacent from vertices in S, the first wave of the snowball sample is given by $S_1 = A(S_0) \cap \bar{S}_0$. The one-wave snowball initiated by S_0 is then given by $S_0 \cup S_1$.

3 A Bernoulli snowball sampling design

Let V_0 be a vertex subset constituting vertices known through some register governed by some action of V_0. In the context of estimating the number of drug users in a city, the register is self-generated by members that are clients at aid agencies.

The unknown process, not controlled by the investigator, performing the selection of the registered clients is modeled by Bernoulli sampling, which means that each client is drawn independently from the population with equal but unknown probability π. Thus, the size of the registered part of the population $v_0|v,\pi \sim \text{bin}(v,\pi)$. Since a frame of V_0 exists, random sampling procedures can be designed exclusively on this set at the initial sampling stage. Like in Tallberg (2003), we let the initial sample S_0 be a simple random sample of size $n = |S_0|$ drawn without replacement from V_0. In the sequel we shall for notational simplicity restrict ourselves to a model with only two blocks.

We shall consider the population V as a realization of some super population, and therefore incorporate a probability denoted by θ, where $z_1, ..., z_v \overset{iid}{\sim}$ Bernoulli (θ) and hence, $v_1 |v,\theta \sim$bin(v,θ).

We will assume that the arcs are independent identically distributed Bernoulli variables with probability 1 on the diagonal (y_{uu}) and with probability β_{z_u,z_w} elsewhere, for $z_u, z_w \in \{1,2\}$. The blocks are labeled so that $\beta_{11} \geqslant \beta_{22}$. Usually we assume $\beta_{11} > \beta_{22}$. If $\beta_{11} = \beta_{22}$, then we label the blocks so that $\beta_{12} > \beta_{21}$. Let $y(S_{0i}, S_{0j}) = r_{ij}$ be the number of nonloop arcs from block i to block j in the initial sample and $y(S_{0i}, S_{1j}) = s_{ij}$ the number of arcs from block i in the initial sample to block j in the first wave of the snowball, where the total number of arcs in S_0, and from S_0 to S_1 are given by $y(S_0, S_0) = r$ and $y(S_0, S_1) = s$, respectively. Hence, conditionally on S_0, each $r_{ij} \sim \text{bin}\left(n_{ij}, \beta_{ij}\right)$ and each $s_{ij} \sim \text{bin}\left\{n_i(v_j - n_j), \beta_{ij}\right\}$, where

$$n_{ij} = \begin{cases} n_i(n_i - 1) & \text{for } i = j \\ n_i n_j & \text{for } i \neq j \end{cases}.$$

For $j = 1, 2$, $k = 0, ..., n_1$ and $l = 0, ..., n_2$, define m_{jkl} to be the number of individuals in $\bar{S}_0 \cap V_j$ that are mentioned by exactly k members of $S_0 \cap V_1$ and l members in $S_0 \cap V_2$. Then

$$\sum_{k=0}^{n_1} \sum_{l=0}^{n_2} m_{jkl} = m_{j00} + m_j,$$

where $m_j = |S_{1j}|$ and $m_{j00} = \left|\bar{S}_0 \cap \bar{S}_1 \cap V_j\right| = v_j - n_j - m_j$. Define further the $(n_1+1)(n_2+1)-1$ vectors

$$\mathbf{m}_j = (m_{j01}, ..., m_{j0n_2}, m_{j10}, ..., m_{j1n_2}, ..., m_{jn_10}, ..., m_{jn_1n_2}), \quad \text{for } j = 1, 2.$$

It is easy to see that $\mathbf{m}_1$ and $\mathbf{m}_2$ are independent and that

$$\mathbf{m}_j \sim \text{multinomial}(v_j - n_j, \mathbf{p}_j),$$

where $\mathbf{p}_j$ is a $\{(n_1+1)(n_2+1)-1\}$-dimensional vector with elements

$$p_{jkl} = \binom{n_1}{k} \beta_{1j}^k \left(1-\beta_{1j}\right)^{n_1-k} \binom{n_2}{l} \beta_{2j}^l \left(1-\beta_{2j}\right)^{n_2-l}.$$

Let $\mathbf{r} = (r_{11}, r_{12}, r_{21}, r_{22})$ be a vector of arc frequencies, $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})$ a vector of arc probabilities and $\mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2)$.

As $\mathbf{r}$ and $\mathbf{m}$ are independent conditional on$\pi, \theta, \boldsymbol{\beta}, v_1$ and v the likelihood function for the case of known blocks can be written

$$\begin{aligned} p(\mathbf{r}, \mathbf{m}, v_0, \mathbf{z} | \pi, \theta, \boldsymbol{\beta}, v_1, v) &= p(\mathbf{r} | \mathbf{m}, v_0, \mathbf{z}, \pi, \theta, \boldsymbol{\beta}, v_1, v)\, p(\mathbf{m} | v_0, \mathbf{z}, \pi, \theta, \boldsymbol{\beta}, v_1, v) \\ &\quad \times p(v_0 | \mathbf{z}, \pi, \theta, \boldsymbol{\beta}, v_1, v)\, p(\mathbf{z} | \pi, \theta, \boldsymbol{\beta}, v_1, v) \\ &= p(\mathbf{r} | v_0, \boldsymbol{\beta}, v_1, v)\, p(\mathbf{m} | v_0, \boldsymbol{\beta}, v_1, v)\, p(v_0 | \pi, v)\, p(\mathbf{z} | v_1, v), \end{aligned}$$

where $\mathbf{z}$ is a $n+m$ vector of known block labels for the vertices in $S_0 \cup S_1$, and we have used that π and θ provides no additional information about $\mathbf{r}$, $\mathbf{m}$ and $\mathbf{z}$ conditional on v_0, v_1 and v. The distribution of $\mathbf{z}$ is computed with the hypergeometric formula as

$$p(\mathbf{z} | v_1, v) = \frac{\binom{v_1}{n_1+m_1}\binom{v_2}{n_2+m_2}}{\binom{v}{n+m}\binom{n+m}{n_1+m_1}} = \frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}}.$$

By inserting the proper conditional distributions given above, the likelihood

is computed as

$$
\begin{aligned}
& p\left(\mathbf{r},\mathbf{m},v_0,\mathbf{z}\,|\pi,\theta,\boldsymbol{\beta},v_1,v\right) \\
= & \prod_{i=1}^{2}\prod_{j=1}^{2}\binom{n_{ij}}{r_{ij}}\beta_{ij}^{r_{ij}}\left(1-\beta_{ij}\right)^{n_{ij}-r_{ij}}\prod_{j=1}^{2}\left[(v_j-n_j)!\prod_{k=0}^{n_1}\prod_{l=0}^{n_2}\frac{p_{jkl}^{m_{jkl}}}{m_{jkl}!}\right] \\
& \times\binom{v}{v_0}\pi^{v_0}\left(1-\pi\right)^{v-v_0}\frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}} \\
= & \left[\prod_{i=1}^{2}\prod_{j=1}^{2}\binom{n_{ij}}{r_{ij}}\beta_{ij}^{r_{ij}}\left(1-\beta_{ij}\right)^{n_{ij}-r_{ij}}\right]\binom{v}{v_0}\pi^{v_0}\left(1-\pi\right)^{v-v_0}\frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}} \\
& \times\prod_{j=1}^{2}\left[\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\left[\prod_{k=1}^{n_1}\prod_{l=1}^{n_2}\frac{\binom{n_1}{k}^{m_{jkl}}\binom{n_2}{l}^{m_{jkl}}}{m_{jkl}!}\frac{}{m_{jkl}!}\right]\prod_{i=1}^{2}\beta_{ij}^{s_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-n_j)-s_{ij}}\right] \\
= & \binom{v}{v_0}\pi^{v_0}\left(1-\pi\right)^{v-v_0}\frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}} \\
& \times\prod_{j=1}^{2}\left[\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\left[\prod_{k=1}^{n_1}\prod_{l=1}^{n_2}\frac{\binom{n_1}{k}^{m_{jkl}}\binom{n_2}{l}^{m_{jkl}}}{m_{jkl}!}\frac{}{m_{jkl}!}\right]\prod_{i=1}^{2}\binom{n_{ij}}{r_{ij}}\beta_{ij}^{t_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-\delta_{ij})-t_{ij}}\right]
\end{aligned}
$$

where $s_{1j}=\sum_{k=0}^{n_1}\sum_{l=0}^{n_2}km_{jkl}$, $s_{2j}=\sum_{k=0}^{n_1}\sum_{l=0}^{n_2}lm_{jkl}$, $t_{ij}=r_{ij}+s_{ij}$ and

$$
\delta_{ij}=\begin{cases}1 \text{ for } i=j\\ 0 \text{ for } i\neq j\end{cases}.
$$

By discarding a multiplicative constant which does not depend on model parameters, the essential part of the likelihood is given by

$$
\begin{aligned}
& f\left(\mathbf{r},\mathbf{m},v_0,\mathbf{z}\,|\pi,\theta,\boldsymbol{\beta},v_1,v\right)=\frac{(v-v_1)\,v_1!}{(v-v_0)!}\pi^{v_0}\left(1-\pi\right)^{v-v_0}\binom{v-n-m}{v_1-n_1-m_1} \\
& \times\prod_{j=1}^{2}\left[\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\prod_{i=1}^{2}\beta_{ij}^{t_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-\delta_{ij})-t_{ij}}\right].
\end{aligned}
$$

A more elaborate and perhaps more interesting case is when only the edge structure y can be observed, i.e. the block labels are unobserved (latent). The probability of observing edge pattern y after discarding a multiplicative

constant which does not depend on model parameters, is then given by

$$g\left(\mathbf{r},\mathbf{m},v_0\left|\pi,\theta,\boldsymbol{\beta},v_1,v,\mathbf{z}\right.\right)=\frac{v!}{(v-v_0)!}\pi^{v_0}\left(1-\pi\right)^{v-v_0}$$
$$\times\prod_{j=1}^{2}\left[\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\prod_{i=1}^{2}\beta_{ij}^{t_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-\delta_{ij})-t_{ij}}\right].$$

4 Prior and posterior distributions

A Bayesian analysis requires the specification of a prior of the parameters $\pi,\theta,\boldsymbol{\beta},v_1$ and v, when the blocks are known, and the computation of the posterior distribution given by

$$p\left(\pi,\theta,\boldsymbol{\beta},v_1,v\left|\mathbf{r},\mathbf{m},v_0,\mathbf{z}\right.\right)\propto p\left(\mathbf{r},\mathbf{m},v_0,\mathbf{z}\left|\pi,\theta,\boldsymbol{\beta},v_1,v\right.\right)p\left(\pi,\theta,\boldsymbol{\beta},v_1,v\right)$$

according to Bayes theorem. The prior distribution can be written as

$$p\left(\pi,\theta,\boldsymbol{\beta},v_1,v\right)=p\left(v_1\left|v,\theta\right.\right)p\left(v\right)p\left(\theta\right)p\left(\boldsymbol{\beta}\right)p\left(\pi\right)$$

if mutual independence is assumed between $\pi,\boldsymbol{\beta}$ and (v_1,v,θ) and between v and θ. The posterior distribution is then given by

$$\begin{aligned}&p\left(\pi,\theta,\boldsymbol{\beta},v_1,v\left|\mathbf{r},\mathbf{m},v_0,\mathbf{z}\right.\right)\\ \propto\;&p\left(\mathbf{r},\mathbf{m},v_0,\mathbf{z}\left|\pi,\theta,\boldsymbol{\beta},v_1,v\right.\right)p\left(v_1\left|v,\theta\right.\right)p\left(v\right)p\left(\theta\right)p\left(\boldsymbol{\beta}\right)p\left(\pi\right)\end{aligned}$$

In the case with unknown blocks, we will arrive at the same posterior distribution although $\mathbf{z}$ is now considered as an unknown parameter instead of observed data.

We assume a priori that $\beta_{ij}\sim$beta(a_{ij},b_{ij}), $\pi\sim$beta(a_π,b_π) and $\theta\sim$beta(a_θ,b_θ). Further, as a discrete informative prior for v which only takes positive values, we choose the zero truncated "discretized" gamma distribution considered in Tallberg (2003). The "discretized" gamma distribution implies that we assign a proportional functional value in a gamma distribution as a probability to the associated discrete outcome; see Bernardo and Smith (1994). As a non-informative alternative prior we consider Jeffreys' (1961) prior given by $p\left(v\right)=1/v$, $v=1,2,...$, which is also the leading term of Rissanen's (1983) prior based on information theory argument.

4.1 With known blocks

In this paper v is the focus of attention, whereas $\pi, \theta, \boldsymbol{\beta}$ and v_1 are regarded as nuisance parameters, or at least parameters of lesser interest. It is therefore our desire to eliminate them from the analysis in order to concentrate on v. In the simpler case with known blocks, this is achieved by computing the marginal posterior distribution of v in the following way

$$\begin{aligned} & p\left(v\left|\mathbf{r}, \mathbf{m}, v_0, \mathbf{z}\right.\right) \propto p\left(\mathbf{r}, \mathbf{m}, v_0, \mathbf{z}\left|v\right.\right) p\left(v\right) \\ = & \sum_{v_1 \in V_1} \int_0^1 \int_0^1 \int_0^1 p\left(\mathbf{r}, \mathbf{m}, v_0, \mathbf{z}\left|\pi, \boldsymbol{\beta}, \boldsymbol{\theta}, v_1, v\right.\right) p\left(\pi\right) p\left(\boldsymbol{\beta}\right) p\left(\theta\right) \\ & \times p\left(v_1\left|v, \theta\right.\right) p\left(v\right) d\pi d\boldsymbol{\beta} d\theta, \end{aligned}$$

By inserting the priors given above, the marginal posterior is computed as

$$\begin{aligned} p\left(v\left|\mathbf{r}, \mathbf{m}, v_0, \mathbf{z}\right.\right) \propto & \sum_{v_1 \in V_1} \frac{(v_1 - n_1)!}{(v_1 - n_1 - m_1)!} \frac{\Gamma\left(n_1\left(v_1 - 1\right) + b_{11} - t_{11}\right)}{\Gamma\left(a_{11} + b_{11} + n_1\left(v_1 - 1\right)\right)} \\ \times & \frac{\Gamma\left(n_2 v_1 + b_{21} - t_{21}\right)}{\Gamma\left(a_{21} + b_{21} + n_2 v_1\right)} \frac{(v - v_1 - n_2)!}{(v - v_1 - n_2 - m_2)!} \frac{\Gamma\left(n_1\left(v - v_1\right) + b_{12} - t_{12}\right)}{\Gamma\left(a_{12} + b_{12} + n_1\left(v - v_1\right)\right)} \\ \times & \frac{\Gamma\left(n_2\left(v - v_1 - 1\right) + b_{22} - t_{22}\right)}{\Gamma\left(a_{22} + b_{22} + n_2\left(v - v_1 - 1\right)\right)} \frac{v!}{(v - v_0)!} \frac{\Gamma\left(v - v_0 + b_\pi\right)}{\Gamma\left(v + a_\pi + b_\pi\right)} \\ \times & \frac{\Gamma\left(v_1 + a_\theta\right) \Gamma\left(v - v_1 + b_\theta\right)}{\Gamma\left(v + a_\theta + b_\theta\right)} \binom{v - n - m}{v_1 - n_1 - m_1} p\left(v\right). \end{aligned}$$

4.2 With unknown blocks

Obviously, when the block structure is unknown an extra component of uncertainty is added to the estimation procedure. With unknown blocks it is difficult to derive an explicit analytical expression for the posterior of v. Since the full conditional posterior distribution of each involved parameter is easy to compute, a feasible approach is to implement the Gibbs sampler algorithm. It is a computer-intensive statistical method which allows us to simulate from the exact posterior distributions. The Gibbs sampler works by iteratively drawing values from each of the full conditional distributions, each conditionally on the last updated values of all the other unknown parameters. As the number of draws approaches infinity, the Gibbs sampler generates accurate samples from the joint posterior distributions. For a more extensive review of the Gibbs sampler algorithm; see Gelman et al (1995) and Gilks,

Richardson and Spiegelhalter (1996). Note that by using Gibbs sampling, we obtain the posterior distributions not only of v but also of the nuisance parameters $\pi, \boldsymbol{\beta}, v_1, \mathbf{z}$ and θ without any additional difficulty. In the case with known blocks, one could implement the Gibbs sampler instead of computing the posterior of v analytically, since both methods will yield the same result, and as a spin-off obtain posterior distributions of $\pi, \boldsymbol{\beta}, v_1$ and θ as well.

- *The full conditional distribution of* v_1

$$\begin{aligned} p\left(v_1 \left| \mathbf{r}, \mathbf{m}, \boldsymbol{\beta}, v, \mathbf{z}, \theta\right.\right) &\propto \frac{v!}{(v-v_0)!}\pi^{v_0}\left(1-\pi\right)^{v-v_0} \\ &\times \binom{v}{v_1}\theta^{v_1}\left(1-\theta\right)^{v-v_1}\frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}} \\ &\times \prod_{j=1}^{2}\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\prod_{i=1}^{2}\beta_{ij}^{t_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-\delta_{ij})-t_{ij}} \\ \propto \left(\frac{\theta}{1-\theta}\right)^{v_1} &\binom{v-n-m}{v_1-n_1-m_1} \\ &\times \frac{(v_1-n_1)!}{(v_1-n_1-m_1)!}\left(1-\beta_{11}\right)^{n_1v_1}\left(1-\beta_{21}\right)^{n_2v_1} \\ &\times \frac{(v-v_1-n_2)!}{(v-v_1-n_2-m_2)!}\left(1-\beta_{12}\right)^{-n_1v_1}\left(1-\beta_{22}\right)^{-n_2v_1}. \end{aligned}$$

- *The full conditional distribution of* v

$$\begin{aligned} p\left(v \left| \mathbf{r}, \mathbf{m}, v_0, \pi, \boldsymbol{\beta}, v_1, \mathbf{z}, \theta\right.\right) &\propto \frac{v!}{(v-v_0)!}\pi^{v_0}\left(1-\pi\right)^{v-v_0} \\ &\times \binom{v}{v_1}\theta^{v_1}\left(1-\theta\right)^{v-v_1}\frac{\binom{v-n-m}{v_1-n_1-m_1}}{\binom{v}{v_1}}p\left(v\right) \\ &\times \prod_{j=1}^{2}\frac{(v_j-n_j)!}{(v_j-n_j-m_j)!}\prod_{i=1}^{2}\beta_{ij}^{t_{ij}}\left(1-\beta_{ij}\right)^{n_i(v_j-\delta_{ij})-t_{ij}} \\ \propto \frac{v!}{(v-v_0)!}\left(1-\pi\right)^{v}\left(1-\theta\right)^{v} &\binom{v-n-m}{v_1-n_1-m_1}p\left(v\right) \\ &\times \frac{(v-v_1-n_2)!}{(v-v_1-n_2-m_2)!}\left(1-\beta_{12}\right)^{n_1v}\left(1-\beta_{22}\right)^{n_2v}. \end{aligned}$$

- *The full conditional distribution of* θ

$$\theta \left| v_1, v \right. \sim \text{beta}\left(v_1 + a_\theta, v + b_\theta - v_1\right).$$

- *The full conditional distribution of* β_{ij}

$$\beta_{ij} \left| \mathbf{r}, \mathbf{m}, v_1, v \right. \sim \text{beta}\left(t_{ij} + a_{ij}, n_i\left(v_j - \delta_{ij}\right) - t_{ij}\right).$$

- *The full conditional distribution of* π

$$\pi \left| v_0, v \right. \sim \text{beta}\left(v_0 + a_\pi, v + b_\pi - v_0\right).$$

- *The full conditional distribution of* $\mathbf{z}$

$$\Pr\left(z_u = 1 \left| \left\{z_w\right\}_{w \neq u}, \mathbf{r}, \mathbf{m}, \boldsymbol{\beta}, v_1, v \right.\right) \propto \binom{v - n - m}{v_1 - n_1 - m_1}$$
$$\times \prod_{j=1}^{2} \frac{(v_j - n_j)!}{(v_j - n_j - m_j)!} \prod_{i=1}^{2} \beta_{ij}^{t_{ij}} \left(1 - \beta_{ij}\right)^{n_i(v_j - \delta_{ij}) - t_{ij}}.$$

5 Concluding remarks

We employ a Bayesian blockmodel approach to estimate the size of small populations with rare properties by using snowball sampling. If contact patterns varies between actors, the estimators can seriously overestimate or underestimate the size of the population, if not taken into account in the model. Therefore we propose a blockmodel approach which ought to improve on the estimators.

Posterior distributions of model parameters are computed when the number of blocks are known. A natural extension would be to generalize the analysis to an unknown number of blocks, and compute the posterior distribution for the number of blocks.

References

[1] Beauchamp, M.A. (1965). An improved index of centrality. *Behavioral Science*, **10**, 161-163.

[2] Bernardo, J.M. and Smith, A.F.M (1994). *Bayesian Theory.* New York: Wiley.

[3] Chow, M. and Thompson, S.K. (1999). Estimation with link-tracing sampling designs - A Bayesian approach. Technical Report 99-03, Department of Statistics, The Penn State University.

[4] Fienberg, S.E. and Wasserman, S. (1981). Categorical data analysis of single sociometric relations. In *Sociological Methodology-1981,* ed. S. Leinhardt San Fransisco: Jossey-Bass, pp. 156-192.

[5] Frank, O. (1979). Estimation of population totals by use of snowball samples. In Holland, P.W. and Leinhardt, S. eds., *Perspectives on Social Network Research.* New York:Academic Press, 319-347.

[6] Frank, O. (2002). Using centrality modeling in network surveys. *Social Networks, **24**,* 385-394.

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

[8] Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry,* **40**, 35-41.

[9] Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, **1**, 215-239.

[10] Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (1995). Bayesian Data Analysis, London: Chapman and Hall.

[11] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo, *in* W. R. Gilks, S. Richardson and D.J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 1-19.

[12] Goodman, L.A. (1961). Snowball Sampling. *Annals of Mathematical Statistics*. **20**, 572-579.

[13] Holland, P., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. *Social Networks*, **5**, 109-137.

[14] Höivik, T. and Gleditsch, N.P. (1975). Structural parameters of graphs: A theoretical investigation. In Blalock, H.M. et al. (eds.). *Quantitative Sociology,* pages 203-223. New York: Academic Press.

[15] Jeffreys, H. (1961). *Theory of probability.* Third Edition. Oxford at the Clarendon Press.

[16] Nieminen, J. (1974). On centrality in a graph. *Scandinavian Journal of Psychology,* **15**, 322-336.

[17] Nowicki, K. and Snijders, T.A.B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association,* **96**, 1077-1087.

[18] Rissanen, J. (1983). A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics.* **11,** No. 2, 416-431.

[19] Snijders, T.A.B. (1981). The degree variance: An index of graph heterogeneity. *Social Networks*, **3**, 163-174.

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

[21] Snijders, T.A.B. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent blockstructure. *Journal of classification,* **14***,* 75-100*.*

[22] Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Methodologie Sociologique.* **36**, 34-58.

[23] Spreen, M. and Coumans, M. (2001). A note on network sampling in drug abuse research. *Connections***. 24(3)**, 44-59.

[24] Tallberg, C. (2003). Estimating the size of hidden populations: A Bayesian approach. Department of Statistics, Stockholm University.

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

[26] Wasserman, S. and Anderson, C. (1987). Stochastic *a posteriori* block-models: Construction and assessment. *Social Networks*, **9**, 1-36.

[27] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: methods and applications*. New York: Cambridge University Press.

[28] White, H.C., Boorman, S. and Breiger, R.L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology,* **81***,* 730-779.