Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Model-based estimation with link-tracing sampling designs

by Steven K. Thompson and Ove Frank

June 2000

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

Statistics Canada Statistique Canada

Canada

Model-based estimation with link-tracing sampling designs

Steven K. Thompson and Ove Frank [1]

Abstract

Samples from hidden and hard-to-access human populations are often obtained by procedures in which social links are followed from one respondent to another. Inference from the sample to the larger population of interest can be affected by the link-tracing design and the type of data it produces. The population with its social network structure can be modeled as a stochastic graph with a joint distribution of node values representing characteristics of individuals and arc indicators representing social relationships between individuals. In this paper maximum likelihood estimators of population graph parameters are described. Predictors of realized population graph quantities are obtained using predictive likelihood. These estimators and predictors are compared with conventional data summaries and illustrated with a numerical example.

Key Words: Snowball samples; Adaptive sampling; Graph sampling; Ignorable designs; Link-tracing designs; Network sampling; Likelihood; Predictive likelihood.

1. Introduction

In studies of hidden and hard-to-access human populations, link-tracing procedures, in which social links are followed from one respondent to another, are commonly involved in obtaining the sample. For example, in a study of injection drug use in relation to the spread of the HIV infection, initial respondents may he asked to identify drug-injection or sexual partners who are then added to the sample. For such a study, the social links are of inherent importance for understanding the issues of interest while at the same time being useful or essential in building the sample. However, inference from the sample to the larger population or social structure of interest can be affected by the link-tracing procedures and the type of data they produce. In this paper we evaluate this inference problem in relation to the design and describe some inference methods for such studies based on maximum likelihood estimation and prediction.

Human populations with social structure are often modeled as graphs, with the nodes of the graph representing individuals and the edges or 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. In real studies of human populations, particularly those that are hidden or hard to access, it is seldom possible to obtain data on the whole population or graph structure. Rather, data are obtained from a sample, and the sample may have been obtained by innovative and unconventional means, including methods taking advantage of the arcs or links from one individual to another. The data may contain information about characteristics of sample individuals, social links within the sample, and in some cases information about links between individuals in the sample and individuals outside the sample.

In this paper we use the term “sampling design” to refer to the procedure by which the sample is selected, whether deliberate or happenstance. For many ethnographic and sociological studies of hidden populations, link-tracing designs are considered the only practical way to obtain a sample large enough to study. In other studies, the social structure is itself the object of interest and the link-tracing methods are used in order to obtain meaningfully structured samples to study.

The statistical literature on design and estimation with link-tracing designs includes procedures variously termed snowball sampling, chain-referral sampling, random walks, nexus sampling, network or multiplicity sampling, and adaptive sampling. A type of link-tracing 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, was termed “snowball sampling” by Goodman (1961). A Bernoulli procedure was assumed for the initial sample. Snowball designs were developed in the graph setting with a variety of initial probability sampling designs and any numbers of links and waves by Frank (1971, 1977a,b, 1978a,b, 1979a), who obtained a variety of design and model based methods for estimating graph quantities from the sample data. Snijders (1992) used the same term “snowball sampling” to include designs in which only a subsample of links from each node is traced. The case in which only one of the links from a node is selected at random and followed to another node, and then one of its links selected, and so on, was called a “random walk” by Klovdahl 1989. Link-tracing sampling methods in which there is only one link from each node have been termed “chains” (Erickson 1979). Frank and Snijders (1994) consider model- and design-based

[1] Steven K. Thompson, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802 USA; Ove Frank, Department of Statistics, Stockholm University, S-10691 Stockholm, Sweden. This research is part of an ongoing, equal collaboration effort and order of authorship was determined by a coin toss.

estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Additional practical and statistical issues in sampling from social networks with various types of snowball, chain-referral, and other link-tracing designs are discussed in Granovetter (1976), Morgan and Rytina (1977), Frank (1979b, 1981, 1988), Watters and Biernacki (1989), van Meter (1990), Spreen (1992), Wasserman and Faust (1994), Spreen and Zwaagstra (1994), Karlberg (1997), Jansson (1997), Spreen (1998), and Robins (1998).

Design-based estimation methods were developed additionally for the closely related designs of network or multiplicity sampling, in which social, kinship, and administrative links were traced (Birnbaum and Sirken 1965, Kalton and Anderson 1986, Levy 1977, Levy and Lemeshow 1991, Sirken 1970, 1972a, b, Sirken and Levy 1974, Sudman, Sirken, and Cowan 1988). For example, in a survey of a rare disease, an initial sample of households might be selected at random and data obtained both for residents of the households and for their siblings. The design-based estimation in these strategies is helped by the symmetry of the links and the encompassing of complete connected components in the sample, and unbiased estimators have been obtained for network sampling with many different initial designs.

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. Following selection of an initial sample of nodes by any of a number of initial designs, the decision on whether to follow links from a node or not depends on the value of a variable of interest observed for the node. For example, in an epidemiological study of a sexually transmitted disease, sexual or social links may be followed only from respondents who have been infected. Design-unbiased estimation methods have been worked out for a wide variety of adaptive cluster sampling strategies.

Design-based methods of inference, such as the design-based estimation methods of network sampling, snowball sampling, and adaptive cluster sampling, 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. The model-based methods described in this paper, on the other hand, do depend on an assumed model for the population or graph. Their practical advantage is that they apply to a wide range of sample selection procedures, and thus allow more leeway in how the sample is actually selected.

In fact many real studies of hidden and hard-to-reach populations use sample selection procedures, including link-tracing, that are not readily analyzed based on idealized design-induced probabilities. For example, in a study to examine the relation of network structure and risk behaviors such as needle sharing among drug injectors in the Bushwick section of Brooklyn, “index” (initial) respondents were used as “auxiliary recruiters” to bring members of their networks into the study (Friedman, Neaigus, Jose, Curtis, Goldstein, Ildefonso, Rothenberg and Des Jarlais 1997, Neaigus, Friedman, Goldstein, Ildefonseo, Curtis and Jose 1995, Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais 1996). Only about 61% of the linked individuals were actually recruited, however. In a long-term study on the heterosexual transmission of HIV infection (Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl 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.

When we compare the maximum likelihood estimators and predictors obtained in this paper with commonly-used conventional estimates or data summaries such as sample means and proportions of node or link values, we find that in most cases the conventional estimates are not the best estimates. Similarly, estimators that would be appropriate if the data included the whole graph may not be appropriate with data on only a sample from the graph. An implication of these results is that 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. The interpretations of this discrepancy provided in this paper give some insight into the conditions under which the best estimate would tend to be lower, or higher, than the conventional one.

Notation and basic issues for design and inference in the graph setting are presented in section 2. In section 3, a wide range of link-tracing procedures, all of which can be analyzed using the approach presented in this paper, are described. In section 4, a class of graph models that we use to illustrate the inference methods of the paper is described. Estimative and predictive maximum likelihood methods for graph parameters and realized population values are described in section 5.

2. Graph models and sampling designs

Consider a graph of N nodes (units) labeled 1, 2, ..., N. Associated with the u^{th} node is a variable of interest Y_u. We denote the full set of node labels $U = \{1, 2, ..., N\}$ and the sequence of node variables by $\mathbf{Y} = (Y_1, ..., Y_N)$. 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 the v^{th} column, is the graph adjacency matrix, denoted $\mathbf{X}$. For convenience we will assume the diagonal elements X_{uu} are zero. The ordered pair (u, v) is sometimes referred to as a dyad of type $(Y_u, Y_v; X_{uv}, X_{vu})$. A graph model is given by a joint probability or density $f(\mathbf{y}, \mathbf{x}; \psi)$ for outcomes $\mathbf{y}$ and $\mathbf{x}$ of $\mathbf{Y}$ and $\mathbf{X}$, respectively, and it may depend on one or more unknown parameters ψ.

A sample s from the graph is a subset of nodes and a subset of node pairs. We can write the combined sample as $s = (s^{(1)}, s^{(2)})$, where $s^{(1)}$ denotes the subset of nodes selected for observation of the associated y-values and $s^{(2)}$ denotes the subset of node pairs selected for observation of the associated x-values. The data consist of the node and node-pair labels in the combined sample together with the associated node and arc-indicator values, that is $d = (u, (v, w), y_u, x_{vw}: u \in s^{(1)}, (v, w) \in s^{(2)})$ or, more simply, $d = (s, \mathbf{y}_{s^{(1)}}, \mathbf{x}_{s^{(2)}})$. Further, it is often convenient to use $\mathbf{y}_s$ to denote the y-values of the nodes in the combined sample and $\mathbf{x}_s$ for the x-values of the node pairs in the combined sample, with $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ denoting the values of the unsampled nodes and node pairs. Often the sampling procedure results in a connection between $s^{(1)}$ and $s^{(2)}$. For example, if all relationships from sample nodes to other sample nodes, and no others, are recorded, then $s^{(2)} = s^{(1)} \times s^{(1)}$. In general, however, the nodes on which y-values are recorded and the node pairs on which x-values are recorded may be quite unrelated sets. In particular, the link-tracing procedures considered in this paper often lead to data on links from nodes in $s^{(1)}$ to nodes outside of $s^{(1)}$.

The sampling design is the procedure by which the sample is selected. This selection procedure may be controlled by the investigators, as is the case with a deliberately implemented probability sampling design, or may be beyond the control of the investigators and determined by the circumstances of the situation. If the probability of selecting the sample does not depend on node values y or link values x or parameters ψ involved in the graph model, we refer to the design as "conventional." For a conventional design the probability of selecting sample s can be written $p(s)$ or $p(s; \varphi)$, where φ denotes any unknown parameters involved in the design (but not the model), as in a Bernoulli sampling with unknown inclusion probability φ for each node. The sampling design may depend on one or more auxiliary variables that are known for the whole population, but that dependence will be left implicit in the notation $p(s)$. Conventional designs include the classical probability designs such as simple random, systematic, stratified, multistage, and unequal probability sampling, as well as model-based purposive and balanced designs based on auxiliary variables.

If the probability of selecting the sample depends on any y or x values, we refer to the design as "adaptive," since the selection procedure adapts to the realized configuration of node and link values in the population. In addition, the design can involve unknown parameters ψ. Thus, in general the sampling design in the graph setting has a selection probability that can be written $p(s \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\psi})$ where $\mathbf{y}$ denotes the sequence of node values, $\mathbf{x}$ the matrix of arc indicator values, and ψ any parameters involved.

Likelihood-based inference, such as maximum likelihood estimation or prediction and Bayes methods, is simplified if the design can be ignored at the inference stage. The fact that the sampling design does not affect the value of a Bayes or likelihood-based estimator in survey sampling was noted by Godambe (1966) for designs that do not depend on any values of the variable of interest and by Basu (1969) for designs that do not depend on values of the variable of interest outside the sample. Scott and Smith (1973) showed that the design could become relevant to inference when the data lacked information about the labels of the units in the sample. Rubin (1976) gave exact conditions for a missing data mechanism – of which a sampling design can be viewed as an example – to be relevant in frequentist and likelihood-based inference. For likelihood-based methods such a maximum likelihood and Bayes methods, the design is "ignorable" if the design or mechanism does not depend on values of the variable of interest outside the sample or on any parameters in the distribution of those values. For frequency-based inference such as design- or model-unbiased estimation, however, the design is relevant if it depends on any values of the variable of interest, even in the sample. Scott (1977) showed that the design is relevant to Bayes estimation if auxiliary information used in the design is not available at the inference stage. Sugden and Smith (1984) gave general and detailed results on when the design is relevant in survey sampling situations. Thompson and Seber (1996) described adaptive designs in which the selection procedure deliberately takes advantage of observed values of the variable of interest, and discussed the relevance of the design in inference from a variety of design and model based perspectives. Similar issues of design and inference arise with adaptive experimental designs, such as medical experiments in which ethical considerations motivate adaptive treatment allocation to favor the more promising treatments as the study progresses (*cf.* Flournoy and Rosenberger 1995, Rosenberger 1996, Wei, Smythe, Lin and Park 1990). It is important to underscore that a design that is said to be "ignorable" for likelihood-based inference might not be ignorable for a frequentist-based inference, such as model-unbiased estimation, and that even though a design may be ignorable at the inference stage, in that for example the way an estimator is calculated does not depend on the design used, the design is still relevant a priori to the properties of the estimator.

The sample data $d = (s, \mathbf{y}_s, \mathbf{x}_s)$ are a function of the sample selected and of the graph values $\mathbf{y}$ and $\mathbf{x}$. The likelihood can be written

$$L(\psi, d) = \sum p(s \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\psi}) f(\mathbf{y}, \mathbf{x}; \boldsymbol{\psi}) \qquad (1)$$

where the sum is over outcomes $(\mathbf{y}, \mathbf{x})$ consistent with the data d. Since the y and x values for nodes and node pairs in the sample are fixed by the data, the sum is over all possible values of the unobserved variables $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ and it actually represents the marginal probability of the sample s selected and the associated observed variables $\mathbf{y}_s$ and $\mathbf{x}_s$.

Thus, in general the likelihood function depends on both the design and the model. The quantity $\sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi)$, based on the model only without consideration of the design, was termed the "face-value likelihood" by Dawid and Dickey (1977) because inference based on this function alone takes the data at face value without considering how the data were selected.

For any design in which the selection of the sample depends on graph y and x values only through those values $\mathbf{y}_s$ and $\mathbf{x}_s$ included in the data, the design probability can be moved out of the sum and forms a separate factor in the likelihood. If in addition the design and model parameters are distinct and not related, the likelihood can be written

$$L(\varphi, \psi, d) = p(s \mid \mathbf{y}_s, \mathbf{x}_s; \varphi) \sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi) \qquad (2)$$

where φ denotes the design parameters and ψ denotes the model parameters. The design then does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators. For any such "ignorable" design, the sum in the above likelihood, over all values of $\mathbf{y}$ and $\mathbf{x}$ leading to the given data value, is simply the marginal probability of the y and x values associated with the sample data. This marginal distribution depends on what sample was selected, but does not depend on how that sample was selected. For likelihood-based inference with a design ignorable in this sense, the face-value likelihood gives the correct inference.

3. Some link-tracing designs

A variety of link-tracing designs are described in this section. Each of these designs is ignorable in the likelihood sense provided the initial sample is selected by an ignorable procedure and provided the data include all the values involved in the selection procedure. Since for all the designs described in this section, the node-pair sample $s^{(2)}$ has a deterministic functional relationship to the node sample $s^{(1)}$, the superscript notation will be omitted and the final node sample $s^{(1)}$ will be denoted simply s.

The simple likelihood methods described in this paper apply to a wide range of ignorable link-tracing designs, including those described in this section. Further research is needed on methods for nonignorable designs, including those with nonignorable selection of the initial sample. Methods for dealing with nonsampling errors such as non-response and reporting errors with link-tracing designs are also in need of further development (*cf.*, Thompson 1997).

3.1 Single-wave design

In a single-wave link-tracing design an initial sample of nodes is selected by any ignorable design from the population of nodes in the graph. For each node in the sample, nodes adjacent from that node are added to the sample. The snowball procedure is assumed to stop after one wave. Thus, node v will be added if for some node u in the initial sample $x_{uv} = 1$.

Let s_0 denote the set of nodes in the initial sample and s_1 denote the added nodes not in the initial sample. 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, $U = s_0 \cup s_1 \cup \bar{s}$. The values $\mathbf{y}$ associated with the nodes can be correspondingly ordered as a sequence $(\mathbf{y}_{s_0}, \mathbf{y}_{s_1}, \mathbf{y}_{\bar{s}})$, where $\mathbf{y}_a = (y_u: u \in a)$ is the subsequence of $\mathbf{y}$ restricted to indices in subset $a \subset U$. The adjacency matrix $\mathbf{x}$ is ordered correspondingly and partitioned into submatrices $\mathbf{x}_{s_0 s_0}$, $\mathbf{x}_{s_0 s_1}$, $\mathbf{x}_{s_0 \bar{s}}$ and so on, where $\mathbf{x}_{ab} = (x_{uv}: u \in a, v \in b)$. Ordering the adjacency matrix in this way facilitates the specification of factors in the likelihood.

With the design above, the probability of selecting sample s depends only on $\mathbf{x}_{s_0 U}$ and so can be written $p(s \mid \mathbf{x}_{s_0 U})$, where $\mathbf{x}_{s_0 U}$ can also be replaced by its column permutation $(\mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 s_1}, \mathbf{x}_{s_0 \bar{s}})$. That is, the probability of selecting the final sample $s = s_0 \cup s_1$ depends on links from the initial sample to other units in the graph, both in s and in $\bar{s}$. The data consist of $(s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$. Since the design does not depend on any x or y values outside the data or on model parameter values, the design is ignorable for likelihood-based inference.

3.2 Multi-wave samples

Consider a snowball sample with $k + 1$ waves after the initial sample. The sample will be denoted $s = s_0 \cup s_1$ with $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \ldots \cup s_{0k}$. An initial sample s_{00} is selected by any design that is ignorable in the likelihood sense. Links are followed and every node with an arc from any node in s_{00} and not already in the sample is added to form the first-wave sample s_{01}. That is, $s_{01} = \{v: x_{uv} = 1$ for some $u \in s_{00}, v \notin s_{00}\}$. Then links are followed in s_{01} to give the second-wave sample $s_{02} = \{v: x_{uv} = 1$ for some $u \in s_{01}, v \notin s_{00} \cup s_{01}\} = \{v: x_{uv} = 1$ for some $u \in s_{00} \cup s_{01}$, $v \notin s_{00} \cup s_{01}\}$. Finally, the $(k+1)$-wave sample, denoted simply s_1, is added by following links from the k^{th} wave sample s_{0k}. That is $s_1 = \{v: x_{uv} = 1$ for some $u \in s_0$, $v \notin s_0\}$. No links from s_1 are followed.

If $s_{0j} = \emptyset$ for any $j < k$ then sampling stops, so that the number of waves added is less than k if at some point there are no links leading out of the current sample to unsampled nodes.

The data consist of sets of node labels in the different waves of the sample and the ordered node pairs from s_0 to U, the sequence of node-values y_s for all nodes in the sample, and the link indicator variables $\mathbf{x}_{s_0 U}$ from s_0 to the set U of nodes in the graph. Thus the data consist of the subgraph data for s_0, that is $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 s_0})$, together with the node values $\mathbf{y}_{s_1}$ for the nodes in the final-wave s_1, the link indicators $\mathbf{x}_{s_0 s_1}$ from s_0 to s_1, and the link indicators $\mathbf{x}_{s_0 \bar{s}}$ from the nodes in s_0 to the nodes not in the sample.

Since the design does not depend on any y or x values outside the data nor on any of the graph model parameters, the design is ignorable and the structure of the data is exactly the same with the $(k+1)$-wave snowball as with the 1-wave snowball design, and with the notation we have used the likelihood and estimation formulas are unchanged with the more general design.

3.3 Completed-wave designs

With a completed snowball sample, the procedure of adding waves is continued until no further links lead out of the sample. Then the number of completed waves K is a random variable and $s_{0,K+1} = s_1$ is the first empty set in the sequence $(s_{00}, s_{01}, \ldots)$. The data are $d = (s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 U})$ or equivalently $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 \bar{s}_0})$. Inference can then proceed with the same likelihood and estimation formulas but with the simplication that the data contains no set s_1 for which $\mathbf{y}_{s_1}$ and $\mathbf{x}_{s_0, s_1}$ are known but from which links are unknown.

3.4 Link-tracing adaptive on node values

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 ($y_u = 1$) and nonusers ($y_u = 0$). 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. Similarly, in a study of sexually transmitted diseases, investigators may be instructed to follow sexual or social links more frequently from infected respondents than from noninfected respondents. The design then can be written $p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})$, since the selection procedure depends on both node and link values. If the data contain all values on which the design depends, that is, $d = (s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})$, then the design is ignorable and maximum likelihood inference is simplified as described in the following sections.

3.5 Tracing only a subsample of sample links

The designs described above can be generalized to procedures in which only a sample of the links leading out from node u in s_0 are followed. Examples include the "random walk" design of Klovdahl (1989) and the generalization of snowball designs described in Snijders (1992). In the random walk design, an initial respondent is asked to give the names of several social contacts. One of these contacts is chosen at random to be interviewed and asked in turn to name several contacts, one of which is chosen at random, and so on. In practice, dead ends can occur when a respondent either reports no contacts or reports only contacts who are already in the sample. In such cases investigators either backtrack and try different leads from previous respondents or find a new initial respondent.

With these subsampling link-tracing designs, the procedure for selecting the sample, though complicated from a design-probability point of view, depends only on values in the sample and on links leading from the sample. We again assume that the initial sample is obtained by any ignorable procedure. Let $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \ldots \cup s_{0k}$ consist of all of the waves from which at least some links are followed. Thus, s_{01} consists of the nodes not previously included obtained by following a subsample of the links from nodes in the initial sample s_{00}, s_{02} consists of the nodes not previously included obtained by following a subsample of the links from nodes in $s_{00} \cup s_{01}$, and so on. No links are followed from the final wave s_1. Allowing for the possibility of dependence on node values, the design can be written $p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})$, so that with data $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, the design is ignorable for likelihood-based inference.

3.6 Data from link-tracing designs

With any of the single or multi-wave link-tracing designs described above, it is of considerable practical importance what data are recorded. If the data include only the sample node labels, the y-values for nodes in the sample, and the arc indicators for pairs of units in the sample, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{ss})$, then the design is nonignorable and must be integrated into the likelihood, which can complicate analysis.

Consider also a study in which social links are used in the design, to find the sample, but only node characteristics (y-values), not relationships are recorded, so that the data are $d = (s, \mathbf{y}_s)$. Then the design is nonignorable.

If on the other hand the data from the link-tracing design include not only the linkages within the sample but the out-linkages (or lack thereof) from all but the last wave to the rest of the graph, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, then the design depends only on graph values in the data and so factors out of the likelihood.

4. A graph model with links related to node values

The likelihood-based approach described in section 2 with sample data from link-tracing designs of types described in section 3 will be illustrated using a class of graph models described in this section. This class of models builds on conditional independence between dyads as in the contact models of Frank (1979a) and Wellman, Frank, Espinoza, Lundquist and Wilson (1991). Conditional on the node values, independence is assumed between dyads, with the distribution of links between pairs of nodes depending on node value. Thus, unconditionally these models have dependence between dyads because of the dependence on the node values. In the models of Holland

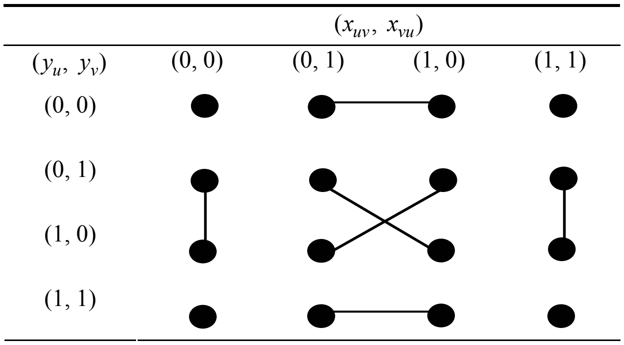
and Leinhardt (1981), dyads are assumed to be independent but with distributions that depend on fixed node parameters. Wasserman (1980) also assumed independence of dyads in modeling the change in a graph over time as a stochastic process. Bayesian extensions and stochasstic blockmodels of Holland, Laskey, and Leinhardt (1983), Fienberg, Meyer, and Wasserman (1985), Wang and Wong (1987), and Frank (1988) provide generalizations to joint distributions with dependence between node values and graph links. Models by Frank and Harary (1982) for randomly colored graphs exhibit a similar structure. The Markov graph models of Frank and Strauss (1986) provide another approach to dependence among dyads but present difficulties for maximum likelihood estimation. Review of a variety of graph models is found in Wasserman and Faust (1994) and Frank (1997).

The maximum likelihood estimation and prediction methods of this paper apply equally to sample data with graph models other than the class of stochastic block models we have used. With other models, the same conditions for ignorability apply. We have chosen this class of models because it is rich enough to encompass important aspects of realism such as dependence between dyads and between dyads and node values, and it is simple enough to have explicit full-graph maximum likelihood estimators for comparison with the estimators based on samples. With other classes of models such as the Markov graph models, estimation even with full-graph data requires numerical methods.

For practical use of the model based approach it is important to have diagnostic tools for evaluations and comparisons between alternative models. For example, with the two-block model used here the conditional independence of dyads could be tested by counting pairs of dyads of different types within and between the blocks. Within each block there are three types of dyads and six types of pairs of dyads. Between the two blocks there are four types of dyads and ten types of pairs of dyads. A Pearson goodness-of-fit statistic between observed and expected counts of the 22 types of pairs of dyads within and between the blocks is asymptotically chi-square distributed with 12 degrees of freedom under the conditional dyad independence assumption. Goodness-of-fit testing for graph models is discussed by Holland and Leinhardt (1981) and Frank and Strauss (1986), and this direction of research needs further development in particular in connection with sample data from link-tracing designs.

In the assumed model the node variables $Y_1, ..., Y_N$ are independent, identically distributed (i.i.d.) Bernoullie random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0, 1$, with $\theta_0 + \theta_1 = 1$. Conditional on the node values $Y_1, ..., Y_N$, the dyads (X_{uv}, X_{vu}) are independent, for $1 \le u < v \le N$, with conditional distribution given by $P[(X_{uv}, X_{vu}) = (k, l) \mid Y_u = i, Y_v = j] = \lambda_{ijkl}$ 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 symmetry requirements are needed $\lambda_{1110} = \lambda_{1101}, \lambda_{1011} = \lambda_{0111}, \lambda_{1010} = \lambda_{0101}, \lambda_{1001} = \lambda_{0110}, \lambda_{0010} = \lambda_{0001}$, and $\lambda_{1000} = \lambda_{0100}$. The pattern of these restrictions is illustrated in Table 1.

Table 1



With these restrictions, it is convenient to introduce the notation

$$\lambda_{ijkl} = \begin{cases} \gamma'_{i+j,k+l,} & \text{if } (ijkl) = (0110) \text{ or } (1001), \\ \gamma_{i+j,k+l,} & \text{otherwise} \end{cases}$$

where $\gamma_{00} + 2\gamma_{01} + \gamma_{02} = 1$, $\gamma_{10} + \gamma_{11} + \gamma'_{11} + \gamma_{12} = 1$, and $\gamma_{20} + 2\gamma_{2521} + \gamma_{22} = 1$. We can interpret γ'_{11} and γ_{11} as the probabilities of dyads with an arc from an unmarked to a marked node only and from a marked to an unmarked node only, respectively. Moreover, for $(ij) \neq (11)$, γ_{ij} is the probability of a dyad with j arcs on i marked and $2 - i$ unmarked nodes.

It will also be convenient to denote $\lambda_{ij1\bullet} = \sum_l \lambda_{ij1l} = \alpha_{ij}$ and $\lambda_{ij11} = \beta_{i+j}$ for $i = 0, 1$ and $j = 0, 1$. Here α_{ij} is the probability of an arc from a node of value i to a node of value j, and β_k is the probability of mutual arcs between k marked nodes.

Let N_i denote the total number of nodes with value i in the graph, for $i = 0, 1$, 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), with $u < v$, such that $(Y_u, Y_v, X_{uv}, X_{vu}) = (ijkl)$.

The likelihood for the full graph under the model with parameters $(\theta, \boldsymbol{\lambda})$ is

$$L(\theta, \boldsymbol{\lambda}; \mathbf{y}, \mathbf{x}) = \left(\prod_{i=0}^{1} \theta_i^{N_i}\right)\left(\prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{k=0}^{1}\prod_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}}\right). \quad (3)$$

In terms of the γs,

$$\prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{k=0}^{1}\prod_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}} = \left(\prod_{i=0}^{2}\prod_{j=0}^{2} \gamma_{ij}^{Rij}\right)(\gamma'^{R'_{11}}_{11})$$

where the Rs are dyad counts corresponding to the pattern in Table 1. That is, $R_{00} = M_{0000}$, $R_{01} = M_{0001} + M_{0010}$,

$R_{02} = M_{0011}$, $R_{10} = M_{0100} + M_{1000}$, $R_{11} = M_{0101} + M_{1010}$, $R'_{11} = M_{0110} + M_{1001}$, $R_{12} = M_{0111} + M_{1011}$, $R_{20} = M_{1100}$, $R_{21} = M_{1101} + M_{1110}$, $R_{22} = M_{1111}$. Note that $R'_{11}(R_{11})$ is the number of dyads with an arc from an unmarked (marked) to a marked (unmarked) node only. Also note that except for $(ij) = (11)$, R_{ij} is the number of dyads on i marked nodes with j arcs.

The maximum likelihood estimators with the whole graph as data are the proportions $\hat{\theta}_i = N_i / N$, $\hat{\gamma}_{ij} = R_{ij} / R_i$, and $\hat{\gamma}'_{11} = R'_{11} / R_1$, where $R_0 = N_0(N_0 - 1)/2$, $R_1 = N_0 N_1$, and $R_2 = N_1(N_1 - 1)/2$. In terms of the λs, this means $\hat{\lambda}_{ijkl} = R'_{11} / R_1$ if $(ijkl) = (0110)$ or (1001) and $\hat{\lambda}_{ijkl} = R_{i+j,k+l} / R_{i+j}$ otherwise.

5. Inference from link-tracing designs

5.1 Estimating graph model parameters

Consider any of the link-tracing designs, for which an initial or multiwave sample 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 data are $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, 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, the likelihood with the sample data given by equation (2) in section 2 is in this case

$$L(\theta, \boldsymbol{\lambda}, d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \sum \left(\prod_{u=1}^{N} \theta_{y_u} \right) \left(\prod_{u<v} \lambda_{y_u y_v x_{uv} x_{vu}} \right)$$

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

Similar to the notation for population counts in the previous section, let $n_i(a)$ denote the number of nodes $u \in a$ with $y_u = i$ for arbitrary subsets $a \subset U$. Let $m_{ijkl}(a, b)$ be the count of pairs of nodes (u, v) such that $u \in a$, $v \in b$, $(y_u, y_v, x_{uv}, x_{vu}) = (ijkl)$, and $u < v$ if both u and v belong to $a \cap b$. An index replaced by a dot means summation over that index. For instance, according to the link-tracing designs described in section 3, only $m_{ijk\bullet}(s_0, s_1)$ is observed, not $m_{ijkl}(s_0, s_1)$.

With data from any of the link-tracing designs described in section 3, the likelihood function is

$$L(\theta, \boldsymbol{\lambda}; d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right)$$
$$\times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \prod_{v \in \bar{s}} \left[\sum_j \theta_j \prod_{ik} \lambda_{ijk\bullet}^{m_{i\bullet k\bullet}(s_0, v)} \right]. \quad (4)$$

For the link-tracing designs in which all links, rather than a subsample, from the initial sample are traced, all of the elements in the submatrix $\mathbf{x}_{s_0 \bar{s}}$ are zero and $m_{i\bullet 0\bullet(s_0, v)} = n_i(s_0)$ for $v \in \bar{s}$, which simplifies the likelihood function to

$$L(\theta, \boldsymbol{\lambda}; d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right)$$
$$\times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})} \cdot \quad (5)$$

The factor $\prod \theta_i^{n_s(s)}$ gives the probability of the observed node values in the sample. The factor $\prod \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)}$ gives the probability of the observed dyad types within $s_0 \times s_0$ given the node values. The factor $\prod \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)}$ gives the probability of the observed dyad types in $s_0 \times s_1$. Since x_{uv} but not x_{vu} is observed, for $u \in s_0$ and $v \in s_1$, the marginal probability that $x_{uv} = k$ given $y_u = i$ and $y_v = j$ is $\lambda_{ijk\bullet}$.

The final factor of (5), with square brackets, gives the probability that there are no arcs from the initial sample to $\bar{s}$. For a node v of the $n(\bar{s})$ nodes outside the sample, θ_j is the probability that $y_v = j$. From any of the $n_i(s_0)$ sample nodes $u \in s_0$ with $y_u = i$, the conditional probability of no link to v, that is, that $x_{uv} = 0$, $\lambda_{ij0\bullet}$.

More formally, the bracketed term can be obtained by conditioning on the number $n_j(\bar{s})$ of nodes of type j in $\bar{s}$. Conditional on $n_j(\bar{s})$, the probability that all the link indicators from s_0 to $\bar{s}$ are zero is obtained as follows. From the $n_i(s_0)$ nodes of type i in s_0 to the $n_j(\bar{s})$ nodes of type j in $\bar{s}$, the probability that all links are zero is $\lambda_{ij0\bullet}^{n_i(s_0) n_j(\bar{s})}$. Using the binomial distribution of $n_1(\bar{s})$ with the law of total probability, the probability that all the links from s_0 to $\bar{s}$ are zero, given y_s, is

$$\sum_{n_1(\bar{s})=0}^{n(\bar{s})} \binom{n(\bar{s})}{n_1(\bar{s})} \left(\prod_j \theta_j^{n_j(\bar{s})} \right) \left(\prod_i \prod_j \lambda_{ij0\bullet}^{n_i(s_0) n_j(\bar{s})} \right)$$
$$= \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})} \cdot (6)$$

With the completed-wave design, the above likelihood expressions are simplified since the terms $m_{ijk\bullet(s_0, s_1)}$ are all zero, so that the factors involving these terms are all equal to one. We also note that $\lambda_{ij0\bullet} = 1 - \alpha_{ij}$ and $\lambda_{ij1\bullet} = \alpha_{ij}$ can be substituted to simplify the likelihood.

5.1.1 Estimative likelihood equations

The maximum likelihood estimators for the parameters θ_1, α_{ij}, and β_k are obtained as the common solutions to the equations

$$\frac{d \log L}{d\theta_1} = \frac{d \log L}{d\alpha_{ij}} = \frac{d \log L}{d\beta_k} = 0 \quad (7)$$

for $i = 0, 1$, $j = 0, 1$, $k = 0, 2$. Differentiating the logarithm of the likelihood (5) with respect to θ_1 and setting equal to zero gives

$$\frac{d \log L}{d\theta_1} = \frac{\partial \log L}{\partial \theta_1} - \frac{\partial \log L}{\partial \theta_0} = 0$$

where the partial derivatives are given by

$$\frac{\partial \log L}{\partial \theta_k} = \frac{n_k(s)}{\theta_k} + n(\bar{s})\frac{\prod_i \lambda_{ik0\bullet}^{n_i(s_0)}}{\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)}}$$

for $k = 0, 1$.

Moreover,

$$\frac{d \log L}{d\alpha_{ij}} = \frac{\partial \log L}{\partial \lambda_{ij10}} + \frac{\partial \log L}{\partial \lambda_{ji01}} - \frac{\partial \log L}{\partial \lambda_{ij00}} - \frac{\partial \log L}{\partial \lambda_{ji00}} \qquad (8)$$

and

$$\frac{d \log L}{d\beta_k} = \sum_{\substack{i,\ j \\ i+j=k}} \left(\frac{\partial \log L}{\partial \lambda_{ij00}} + \frac{\partial \log L}{\partial \lambda_{ij11}} - \frac{\partial \log L}{\partial \lambda_{ij01}} - \frac{\partial \log L}{\partial \lambda_{ij10}} \right) \qquad (9)$$

where the partial derivatives are given by

$$\frac{\partial \log L}{\partial \lambda_{ijkl}} = \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijk\bullet}(s_0, s_1)}{\lambda_{ijk\bullet}} + (1-k)\,n(\bar{s})\frac{\theta_j\, n_i(s_0)\, \lambda_{ij0\bullet}^{n_i(s_0)-1}}{\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)}}.$$

It is convenient to write the likelihood equation for θ_1 as

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} + \frac{n(\bar{s})(\rho - 1)}{\theta_1 \rho + \theta_0} = 0 \qquad (10)$$

where

$$\rho = \prod_{i=0}^{1} \left(\frac{\lambda_{i10\bullet}}{\lambda_{i00\bullet}} \right)^{n_i(s_0)} = \prod_{i=0}^{1} \left(\frac{1 - \alpha_{i1}}{1 - \alpha_{i0}} \right)^{n_i(s_0)}.$$

Note that $\rho = \rho_0^{n_0(s_0)} \rho_1^{n_1(s_0)}$, where $\rho_i = (1 - \alpha_{i1})/(1 - \alpha_{i0})$ is the ratio between the probabilities of no arc from an i-node to a positive and a zero node, respectively.

An interpretation of the influence of the graph structure on estimation of θ_1 is provided by considering the graph parameters α – and hence ρ – as fixed. Denote the sample proportion of positive nodes by $\hat{\theta}_c = n_1(s)/n(s)$. This is the conventional or naive estimator of θ_1, using the sample proportion of positive nodes. If $\rho = 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be $\hat{\theta}_c$. If $\rho < 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be less than $\hat{\theta}_c$, and if $\rho > 1$, $\hat{\theta}_1 > \hat{\theta}_c$. In particular, $\alpha_{i1} = \alpha_{i0}$ for $i = 0, 1$ implies $\rho = 1$ and the maximum likelihood estimator is $\hat{\theta}_1 = \hat{\theta}_c$.

Consider for instance the case in which for any given value of y_u, a link from node u to node v is more likely when $y_v = 1$ than when $y_v = 0$, so that $\alpha_{i1} > \alpha_{i0}$, for $i = 0, 1$. Then $(1 - \alpha_{i1})/(1 - \alpha_{10}) < 1$, for $i = 0, 1$, so that $\rho < 1$ and the maximum likelihood estimator $\hat{\theta}_1$ is less than the conventional estimator $\hat{\theta}_c$. One could say that the link-tracing design is leading investigators to an unrepresentatively high proportion of positive nodes, and the maximum likelihood estimator is adjusting for this.

In specific cases some of the λ_{ijkl} might be set to zero and the likelihood equations have to be modified accordingly. Some specific cases will now be illustrated.

5.1.2 A symmetric model

Symmetric models have $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always mutual 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 $\lambda_{ij00} + \lambda_{ij11} = 1$. Here $\lambda_{ij11} = \beta_{i+j} = \alpha_{ij} = \alpha_{ji}$ and

$$\rho = \prod_{i=0}^{1} \left(\frac{1 - \beta_{i+1}}{1 - \beta_i} \right)^{n_i(s_0)}.$$

Letting $m_{ijkl}(s_0, s) = r_{i+j,k+l}$, we obtain the maximum likelihood estimators as the solutions to the equations

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} + \frac{n(\bar{s})(1 - \rho)}{\theta_0 + \rho\theta_1} = 0 \qquad (11)$$

$$\frac{r_{02}}{\beta_0} - \frac{r_{00}}{1 - \beta_0} - \frac{n(\bar{s})\, n_0(s_0)\, \theta_0}{(1 - \beta_0)(\theta_0 + \rho\theta_1)} = 0 \qquad (12)$$

$$\frac{r_{12}}{\beta_1} - \frac{r_{10}}{1 - \beta_2} - \frac{n(\bar{s})[n_1(s_0)\theta_0 + n_0(s_0)\rho\theta_1]}{(1 - \beta_1)(\theta_0 + \rho\theta_1)} = 0 \qquad (13)$$

$$\frac{r_{22}}{\beta_2} - \frac{r_{20}}{1 - \beta_2} - \frac{n(\bar{s})\, n_1(s_0)\, \rho\theta_1}{(1 - \beta_2)(\theta_0 + \rho\theta_1)} = 0. \qquad (14)$$

If the symmetric model is further simplified by assuming $\beta_0 = \beta_1 = 0$, there are only the two parameters θ_1 and β_2, and the equations to be solved are

$$\theta_1 \beta_2 = r_{22} / N n_1(s_0)$$

and

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1 - \beta_2)^{n_1(s_0)}.$$

For instance suppose the value $y_u = 1$ indicates injection drug use and $x_{uv} = 1$ indicates u and v are injection partners, so that links are only possible between users and tracing these links can only add users to the sample. As an illustration, consider a population of size N = 10,000 with statistics $n_1(s_0) = 7, n_0(s_0) = 43, n_1(s) = 47,$ and $r_{22} = 42$. The likelihood equations are $\theta_1 \beta_2 = 0.0006$ and $(10{,}000 - 47/\theta_1)/10{,}000 - 43/\theta_0) = (1 - \beta_2)^7$, leading to the maximum likelihood estimators $\hat{\theta}_1 = 0.12$ and $\hat{\beta}_2 = 0.005$. The naive estimator for θ_1 in this case would be the sample proportion 47/90 = 0.52 and the naive estimator for β_2 would be

$$42\Big/\binom{47}{2} = 0.039,$$

the proportion of links between users in the sample out of the number possible.

5.1.3 An asymmetric model

A specific asymmetric model has $\lambda_{ijkl} = \lambda_{ijk\bullet}\lambda_{ij\bullet l} = \lambda_{ijk\bullet}\lambda_{jil\bullet}$, so that all arcs are independent. Now $\beta_{i+j} = \alpha_{ij}\alpha_{ji}$ and we obtain the maximum likelihood estimators as the solutions to the equations

$$\rho = \frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0}$$

and

$$\frac{\alpha_{ij}}{1-\alpha_{ij}} = \frac{m_{ij1}}{m_{ij0} + n_i(s_0)\rho^j\,\theta_j\,(N - n_0(s)/\theta_0)}$$

for $i = 0, 1$ $j = 0, 1$, where $m_{ijk} = m_{ijk\bullet}(s_0, s)$.

In particular, if we specify this asymmetric model by $\alpha_{ij} = ij\alpha$, so that arcs are possible with probability α only between marked nodes, then the equations to be solved are

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1-\alpha)^{n_1(s_0)}$$

and

$$\frac{\alpha}{1-\alpha} = \frac{m_{111}}{m_{110} + [N\theta_1 - n_1(s)]n_1(s_0)}.$$

Again, iterative methods are appropriate.

5.2 Predictive likelihood for the total of the unobserved node values

For predicting the value of the unobserved random variable $n_1(\bar{s})$ from the observed data, the relevant likelihood is

$$L[\theta, \boldsymbol{\lambda}; d, n_1(\bar{s})] = p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0U})$$
$$\times\left(\prod_i \theta_i^{n_i(s)+n_i(\bar{s})}\right)\binom{n(\bar{s})}{N_1(\bar{s})}\left(\prod_{ijkl}\lambda_{ijkl}^{m_{ijkl}(s_0,\, s_0)}\right)$$
$$\times\left(P\prod_{ijk}\lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)}\right)\left(\prod_{ij}\lambda_{ij0\bullet}^{n_i(s_0)n_j(\bar{s})}\right). \qquad (15)$$

Use of the term “prediction” implies only that the object of inference is a random variable rather than a fixed, unknown parameter, and does not necessarily imply forecasting in time.

The estimative likelihood for $n_1(\bar{s})$ is obtained from (15) by substituting the estimates $\hat{\theta}$ and $\hat{\lambda}$ that maximize the (marginal) likelihood (5). The value of $n_1(\bar{s})$ maximizing the estimative likelihood would be the estimative maximum likelihood predictor of $n_1(\bar{s})$. While estimative likelihood methods tend to produce reasonable point predictions in many cases, they are less useful as a basis for prediction intervals, since the estimates of the parameters are in essence treated as the true values (*cf.*, Bjørnstad 1990, 1996, Lejeune and Faulkenberry 1982). For this reason, we emphasize the use of the profile predictive likelihood.

Rather than substituting fixed estimators of the parameters into (15) and maximizing this estimative likelihood with respect to $n_1(\bar{s})$, the likelihood (15) is now simultaneously maximized with respect to both parameters and $n_1(\bar{s})$. This means that for each value of $n_1(\bar{s})$ there are parameter values $\tilde{\theta}_i\,[n_1(\tilde{s})]$ and $\tilde{\lambda}_{ijkl}\,[(n_1(\bar{s})]$ which maximize (15) with respect to θ and λ. Substitution of these values into (15) defines the profile likelihood $L_p[n_1(\bar{s}); d]$ for $n_1(\bar{s})$. The value of $n_1(\bar{s})$ maximizing the profile likelihood is the profile maximum likelihood predictor of $n_1(\bar{s})$.

For any given value of $n_1(\bar{s})$, the likelihood is maximized where the derivatives with respect to the remaining parameters equal zero. The maximizing values of θ_i are straightforward and are given by

$$\tilde{\theta}_i = \frac{n_i(s) + n_i(\bar{s})}{N}. \qquad (16)$$

For the remaining parameters we use $d\ \log L/d\alpha_{ij}$ and $d\ \log L/d\beta_k$ from (8) and (9), with the partial derivatives now given by

$$\frac{\partial \log L}{\partial\lambda_{ijkl}} = \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijk\bullet}(s_0, s_1)}{\lambda_{ijk\bullet}}$$
$$+ (1-k)\frac{n_i(s_0)\,n_j(\bar{s})}{\lambda_{ij0\bullet}} \qquad (17)$$

Note that the $n_j(\bar{s})$ for $j = 0,\ 1$ are contained in (15) only in the factors

$$\binom{n(\bar{s})}{n_1(\bar{s})}\prod_j \Lambda_j^{n_j(\bar{s})}$$

where $\Lambda_j = \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)}$. Since L is proportional to a binomial probability with parameters $n(\bar{s})$ and $\Lambda_1/(\Lambda_0 + \Lambda_1)$, it follows that the maximum of L over $n_1(\bar{s})$ is obtained for $n_1(\bar{s})$ equal to the integer closest to

$$\frac{n(\bar{s})\Lambda_1}{\Lambda_0 + \Lambda_1} + \frac{\Lambda_1 - \Lambda_0}{2(\Lambda_0 + \Lambda_1)}$$

or either of the integers closest to this number if there are two of them. In fact (see, for instance, Feller 1957, page 140), the mode of a binomial distribution with parameters $(n,\ p)$ is the integer in the interval $[(n+1)p - 1, (n+1)p]$ or either of the endpoints if they are integers. Thus, the mode is the integer or the integers that are closest to the interval midpoint $(n+1)p - (1/2) = np + (p-q)/2$, where $q = 1 - p$.

If initial values of the parameter estimators are obtained from the solution of (7) and substituted into the Λ_j, then a predicted value $n_1(\bar{s})$ is given as above. If this predicted value is inserted into (16) and (17), then new estimates of the parameters are obtained that can be substituted into the Λ_j to find a new predicted value of $n_1(\bar{s})$, continuing until the values converge to the solution minimizing (15). Alternatively, the solution can be found by direct computation of the likelihood (15) for different values of $n_1(\bar{s})$, substituting the solutions obtained from (16) and (17) for the parameter values.

5.2.1 Example: Symmetric model

The predictive likelihood equation (15) for the symmetric model is

$$L[\theta,\beta;d,n_1(\bar{s})]=p(s|\mathbf{y}_s,\mathbf{x}_{s_0U})\left(\prod_i \theta_i^{n_i(s)+n_i(\bar{s})}\right)\binom{n(\bar{s})}{n_1(\bar{s})}$$

$$\times\left(\prod_{i,\,j}\beta_{i+j}^{m_{ij11(s_0,s)}}(1-\beta_{i+j})^{m_{ij00}(s_0,s)+n_i(s_0)n_j(\bar{s})}\right)\cdot(18)$$

Let $r_{kl} = r_{kl}(s_0, s)$ denote the count of node pairs in $s_0 \times s$ with total node value k and total number of links l. With the symmetric model, l can take only the values 0, indicating no link between the nodes, or 2, indicating a symmetric link. In particular, $r_{02} = m_{0011}(s_0, s)$, $r_{12} = m_{0111}(s_0, s) + m_{1011}(s_0, s)$, and $r_{22} = m_{1111}(s_0, s)$ denote the sample counts of links between nodes of total value k, for $k = 0, 1, 2,$ respectively. With this notation the last factor in (18) can be written

$$\prod_{k=0}^{2}\beta_k^{r_{k2}}(1-\beta_k)^{r_{k0}+\sum_{\substack{i,\,j\\ i+j=k}} n_i(s_0)n_j(\bar{s})}.$$

Denote by $c_k = c_k[n_1(\bar{s})]$ the number of possible node pairs in $s_0 \times U$ having total value k, so that

$$c_k = r_{k\bullet} + \sum_{\substack{i,\,j\\ i+j=k}} n_i(s_0)n_j(\bar{s})$$

$$= \sum_{\substack{i,\,j\\ i+j=k}} n_i(s_0)\left[n_j(s) + n_j(\bar{s})\right].$$

For any given value of $n_1(\bar{s})$, the likelihood is maximized by $\tilde{\theta}_i = [n_i(s) + n_i(\bar{s})]/N$ for $i = 0,\ 1$ and $\tilde{\beta}_k = r_{k2}/c_k$ for $k = 0, 1, 2$. Note that $\tilde{\theta}$ and the $\tilde{\beta}_k$ are functions of the unobserved variable $n_1(\bar{s})$.

The profile predictive likelihood function for $n_1(\bar{s})$ is obtained by substituting the maximizing values $\tilde{\theta}$ and $\tilde{\beta}_k$ for the parameters in (18), giving

$$L_p[n_1(\bar{s});\ d]=p(s|\mathbf{y}_s,\mathbf{x}_{s_0U})\left(\prod_i\left(\frac{n_i(s)+n_i(\bar{s})}{N}\right)^{n_i(s)+n_i(\bar{s})}\right)$$

$$\times\binom{n(\bar{s})}{n_1(\bar{s})}\left(\prod_k\left(\frac{r_{k2}}{c_k}\right)^{r_{k2}}\left(1-\frac{r_{k2}}{c_k}\right)^{c_k-r_{k2}}\right)$$

which is a function of $n_1(\bar{s})$ alone. The maximum profile likelihood predictor of $n_1(\bar{s})$, easily obtained by straightforward computation, is an integer between 0 and $n(\bar{s})$ giving the largest value of (19).

5.3 On assessing accuracy of estimates

For confidence intervals and other forms of inference, the inverse of the observed Fisher information $\mathbf{I}(\hat{\varphi})$ is suggested, where $\hat{\varphi}$ is the vector of parameter maximum likelihood estimates and $\mathbf{I}$ is the matrix of negated second derivatives of the log likelihood function evaluated at those estimated values. The use of the observed, as opposed to expected, Fisher information to assess the accuracy of an estimate is described in Efron and Hinkley (1978). More recently, Lindsay and Li (1997) argue that the observed information gives a better assessment of the realized, as opposed to expected, error of the estimate. In developing large-sample approximations to the properties of the estimators of θ and λ it is important to make appropriate assumptions about how λ depends on N so that the graph model and the sample do not degenerate. See for instance the asymptotic results for some simple graph models given by Palmer (1985).

As with the calculation of the maximum likelihood estimates themselves, the calculation of the observed information matrix is not affected by the link-tracing sampling design, since the design is ignorable for likelihood based on inference. This is in contrast to the expected Fisher information, the value of which is affected by the design in addition to the graph model, unless the design is a conventional one not depending on any $\mathbf{y}$ and $\mathbf{x}$ values.

For a $(1-\varepsilon)$-level prediction interval for a random variable such as $n_1(\bar{s})$, one method would be to use a central region having mass $(1-\varepsilon)$ of the normalized profile likelihood function for $n_1(\bar{s})$ (*cf.*, Bjørnstad 1990, 1996). For the symmetric model, the $(1-\varepsilon)$ prediction interval for $n_1(\bar{s})$, is readily obtained by computing (19) for $n_1(\bar{s}) = 0$, 1, 2, …, until the computed values become negligible, normalizing by dividing by the cumulative total $\sum_{n_1(\bar{s})=0}^{n(\bar{s})} L_p$ and using the $\varepsilon/2$ and $1-\varepsilon/2$ quantiles as the interval endpoints.

Acknowledgements

Support for this research was provided by the National Science Foundation (DMS-9626102), the National Institutes of Health, National Institute on Drug Abuse (R01DA 09872), and the Swedish Council for Research in the Humanities and the Social Sciences (HSFR F 0750/96).

References

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā* A, 31, 441-454.

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*, 2, 11. Washington: Government Printing Office.

Bjørnstad, J.F. (1990). Predictive likelihood: A review. *Statistical Science*, 5, 242-265.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

Dawid, A.P., and Dickey, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72, 845-850.

Efron, B., and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika*, 65, 457-487.

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

Fienberg, S.E., Meyer, M.M. and Wasserman, S.S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80, 51-67.

Flournoy, N., and Rosenberger, W.F., Eds. (1995). *Adaptive Designs*. Hayward, CA: Institute of Mathematical Statistics.

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

Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

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

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

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

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

Frank, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciencies*, 319, 207-218.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 110-155.

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

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

Frank, O., and Harary, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77, 835-840.

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

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

Friedman, S.R., Neaigus, A., Jose, B., Curtis, R., Goldstein, M., Ildefonso, G., Rothenberg, R.B. and Des Jarlais, D.C. (1997). Sociometric risk networks and HIV risk. *American Journal of Public Health*. In press.

Godambe, V.P. (1966). A new approach to sampling from finite populations. I. *Journal of the Royal Statistical Society* B, 28, 310-319.

Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.

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

Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic block-models: First steps. *Social Networks*, 5, 109-137.

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

Jansson, I. (1997). On statistical modeling of social networks. Ph.D. Thesis, Stockholm University.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society* A, 149, 65-82.

Karlberg, M. (1997). Triad count estimation and transitivity testing in graphs and digraphs. Ph.D. Thesis, Stockholm University.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In, *The Small World*, (Ed. M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.

Lejeune, M., and Faulkenberry, G.D. (1982). A simple predictive density function. *Journal of the American Statistical Association*, 77, 654-657.

Levy, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association*, 72, 758-763.

Levy, P.S., and Lemeshow, S. (1991). *Sampling of Populations*; *Methods and Applications*. New York: John Wiley & Sons, Inc.

Lindsay, B.G., and Li., B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics*, 25, 2172-2199.

Morgan, D.L., and Rytina, S. (1977). Comment on “Network sampling: some first steps” by Mark Granovetter. *American Journal of Sociology*, 83, 722-727.

Neaigus, A., Friedman, S.R., Goldstein, M.F., Ildefonseo, 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*, (Eds., R.H. Needle, S.G. Genser and R.T. II Trotter). 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.

Palmer, E.M. (1985). *Graphical Evolution*. New York: John Wiley & Sons, Inc.

Robins, G.L. (1998). Personal attributes in inter-personal contexts: Statistical models for individual characteristics and social relationships. Ph.D. Thesis, University of Melbourne.

Rosenberger, W.F. (1996). New directions in adaptive designs. *Statistical Science*, 11, 137-149.

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 Transmission*, (Eds., R.H. Needle, S.G. Genser and R.T. II Trotter). 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.

Scott, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā* C, 39, 1-9.

Scott, A.J., and Smith, T.M.F. (1973). Survey design, symmetry, and posterior distributions. *Journal of the Royal Statistical Society* B, 35, 57-60.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.

Sirken, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.

Sirken, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.

Sirken, M.G., and Levy, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, 69, 68-73.

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

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

Spreen, M. (1998). Sampling personal network structures: Statistical inference in ego-graphs. Ph.D. Thesis, University of Groningen.

Spreen, M., and Zwaagstra, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: Introducing the network concept in studies of hidden populations. *International Sociology*, 9, 475-491.

Sudman, S., Sirken, M.G. and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.

Sugden, R.A., and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

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

Thompson, S.K. (1997). Adaptive sampling in behavioral surveys. In *The Validity of Self-Reported Drug Use*: *Improving the Accuracy of Survey Estimates*, (Eds., L. Harrison and A. Hughes). NIDA Research Monograph 167, Rockville, MD: National Institute of Drug Abuse, 296-319.

Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

van Meter, K.M. (1990). Methodological and design issues: techniques for assessing the representatives of snowball samples. In *The Collection and Interpretation of Data from Hidden Populations*, (Ed., E.Y. Lambert). NIDA Monograph 98. Rockville, MD: National Institute on Drug Abuse, 31-43.

Wang, Y.J., and Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8-19.

Wasserman, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75, 280-294.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis*: *Methods and Applications*. New York: Cambridge University Press.

Watters, J.K., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

Wei, L.J., Smythe, R.T., Lin, D.Y. and Park, T.S. (1990). Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association*, 85, 156-162.

Wellman, B., Frank, O., Espinoza, V., Lundquist, S. and Wilson, C. (1991). Integrating individual, relational and structural analysis. *Social Networks*, 13, 223-249.