Network Surveys

Ove Frank

This review presents some general concepts and approaches that are used in network surveys, and it also tries to give a flavor of what special methodological issues and modern interesting applications can be encountered in this field. To this end, recursive calculations of inclusion probabilities in successive snowball samples are discussed. Moreover, investigations of the World Wide Web and of criminal co-offender networks are mentioned as examples of applications that require special network methodology.

1. Introduction

It is a pleasure to contribute to this volume honoring Daniel Thorburn and be able to refer to his interest in network surveys as shown, for instance, in his contribution (Thorburn 2002) to my own festschrift. There he illustrates how network data from a one-wave snowball sample can be used to develop generalized difference estimators.

The Statistics Department of Stockholm University has a long tradition of research in survey sampling methodology. Network surveys belong to the more recent part of this tradition. Several teachers and students at our department have contributed to this development of modern statistical methodology.

A comprehensive Encyclopedia of Complexity and Systems Science edited by Meyers (2009) and published in eleven volumes by Springer is a huge reference work that gives extensive demonstrations of the use of network models and network surveys in various fields. In Frank (2005, 2009, 2010) more elaborated treatments are given of the main models and methods used in network surveys. An excellent recent textbook focusing on statistical issues in network surveys is Kolaczyk (2009), and it is strongly recommended for anyone who might consider penetrating this challenging field.

The present review is by necessity brief and subjective and reflects my interest in specific topics in a vivid and accelerated development of models and applications. Basic concepts of graphs and networks are defined in the next section, and representations of variables and data are discussed. The two following sections present some common network models and basic sampling procedures with special attention paid to snowball sampling.

2. Graphs and Networks

A graph (digraph) G on vertex set V and edge set E is generally defined as a function that associates with each edge an unordered (ordered) pair of vertices. If the function is a bijection, so that different edges correspond to different pairs of vertices, the graph is said to be simple, and the edge set can be considered as a subset of V^2. A simple graph or

1 Department of Statistics, Stockholm University, SE-10691 Stockholm, Sweden. Email: ove.frank@telia.com
digraph can be represented by an adjacency function

\[ y = \{(u, v, y(u, v)) : (u, v) \in V^2\} \]

defined on \( V^2 \) with binary values indicating whether or not vertex \( u \) has an edge to vertex \( v \), that is

\[ y(u, v) = 1 \text{ if } G^{-1}(u, v) \in E \]
\[ y(u, v) = 0 \text{ otherwise.} \]

If \( V = \{1, \ldots, N\} \), the values of the adjacency function \( y \) can be displayed as an adjacency matrix with \( N \) rows and \( N \) columns.

If one or more categorical variables are defined on the edges, the graph can be drawn with colored edges representing different categories of the different edge-variables. For instance, consider two edge-variables representing friendship and co-operation between people and assume that there are three nonnull strength-categories weak, medium and strong of each kind of relationship. This would require at most fifteen edge colors in a simple graph, or, equivalently, at most sixteen categories of vertex pairs corresponding to two categorical variables with four categories on each.

Graphs with categorical or numerical variables specifying attributes or properties of the edges are called valued graphs or networks. The adjacency function defined as an edge-indicator on the vertex pairs in a simple graph is generalized to a categorical or numerical edge-variable in a network. There can be several such variables corresponding to edges of different relationships, and they can be simultaneously represented by a multivariate function \( y \) defined on the set of vertex pairs. In addition to the edge-variables in \( y \) there might be attributes or properties specified for the vertices, so that there is also a univariate or multivariate function

\[ x = \{(v, x(v)) : v \in V\} \]

defined on the vertex set. The variables in \( x \) and \( y \) allow for a rich variety of network structures with interdependencies between the vertex- and edge-variables.

As a simple illustration, consider a school class of \( N \) children with a variable \( x = (x_1, x_2) \) giving gender \( x_1 \) and achievement \( x_2 \) and a variable \( y \) giving mutual friendship estimated as null, weak or strong. Gender has two categories and achievement five, say. It follows that the \( N \) children are distributed over ten possible individual categories, and the \( N(N - 1)/2 \) unordered pairs of children are distributed over three friendship categories. The numbers of null, weak and strong friendships within and between the ten possible individual categories give us 165 possible frequencies. How should they be analyzed to tell us about the interdependencies between gender, achievement and friendship? Could we decide whether individual homogeneity increases friendship, or whether friendship has a positive or negative influence on achievement? How should data be selected, combined and analyzed if they are available during several semesters from different school classes and different schools? Obviously there are many statistical issues for network data that might not be properly analyzed if the network aspect is discarded.
3. Random Networks

Let $Y$ be a random variable with outcomes $y$ that represent networks in some specified family $F$ of networks. The probability distribution of $Y$ induces a probability distribution over $F$ that represents a random network model. Especially for large networks, a detailed specification of the properties or characteristics of an observed network $y$ is impractical, if not impossible, and of less use than some appropriately selected summary statistics that might catch the general tendencies or patterns governing the creation of the network. This is the reason why a study of random network models might be beneficial for understanding complex networks.

Even very simple network models can be hard to analyze and can reveal surprising structural properties. A good example is the classical random graphs introduced by Erdős and Renyi half a century ago. The uniform random graph with $n$ vertices and $m$ edges uniformly selected among the unordered vertex pairs has totally dominated the random graph theory literature for decades. The model is a uniform distribution over family $F$ of all graphs on $V = \{1, \ldots, n\}$ with $m$ edges. $F$ consists of $N(m)/m!$ graphs, where $N = n(n - 1)/2$. Various limiting results on connectivity, clustering, and other properties have been proved for increasing values of $n$ and $m$ and with specific relations between them. Similar results have also been achieved for the Bernoulli ($V, p$) graph defined on vertex set $V = \{1, \ldots, n\}$ with edge indicators $Y_{uv}$ that are Bernoulli($p$) distributed and independent for $1 \leq u < v \leq n$. Here family $F$ consists of all the $2^N$ graphs on $V$ with probability $p^m(1 - p)^{N - m}$ attached to any of the $N(m)/m!$ graphs with $m$ edges for $m = 0, 1, \ldots, N$ and $N = n(n - 1)/2$.

An important class of network models used in the social sciences is given by random digraphs on $V = \{1, \ldots, n\}$ with independent dyads $(Y_{uv}, Y_{vu})$ for $1 \leq u < v \leq n$. Such digraphs have $3n(n - 1)/2$ degrees of freedom corresponding to the probabilities

$$P(Y_{uv} = i, Y_{vu} = j) = P_{uvij}$$

for $i = 0, 1$ and $j = 0, 1$ satisfying

$$\sum_i \sum_j P_{uvij} = 1$$

for $1 \leq u < v \leq n$. By making special assumptions about these probabilities, the degrees of freedom can be reduced to $3n - 1$ corresponding to vertex parameters $(\alpha_v, \beta_v, \gamma_v)$ governing the numbers of out-edges, in-edges and mutual edges at vertex $v$ according to

$$P_{uvij} = \exp[-\lambda_{uv} - i(\alpha_u + \beta_v) - j(\alpha_v + \beta_u) - ij(\gamma_u + \gamma_v)].$$

Here $\lambda_{uv}$ is determined by requiring the probabilities to sum to unity for each $(u, v)$. The probability is invariant if a common constant is added to all $\alpha_v$ and subtracted from all $\beta_v$, which implies that the $3n$ vertex parameters satisfy a linear restriction $\sum_v \alpha_v = \sum_v \beta_v$. 
It is convenient to set the two sums to zero and introduce a compensating parameter \( \mu \) for edges. With the out- and in-effects centered at zero the model is

\[
P_{uvij} = \exp[-\lambda_{uv} - i(\mu + \alpha_u + \beta_v) - j(\mu + \alpha_v + \beta_u) - ij(\gamma_u + \gamma_v)],
\]

and the likelihood

\[
L = P(Y_{uv} = y_{uv} \text{ for } 1 \leq u < v \leq n)
\]

is obtained as

\[
L = \exp \sum_{u<v} \left[ -\lambda_{uv} - y_{uv}(\mu + \alpha_u + \beta_v) - y_{vu}(\mu + \alpha_v + \beta_u) - y_{uv}y_{vu}(\gamma_u + \gamma_v) \right]
\]

\[
= \exp[-\lambda - \mu y_v - \sum_v \alpha_v y_v - \sum_v \beta_v y_v - \sum_v \gamma_v \sum_u y_{uv}y_{vu}]
\]

where

\[
y_{vv} = 0
\]

\[
y_v = \sum_{u \in V} y_{vu}
\]

\[
y_v = \sum_{u \in V} y_{uv}
\]

\[
y_v = \sum_{u \in V} \sum_{v \in V} y_{uv}
\]

and \( \lambda \) is the sum of the normalizing constants. Thus the log-likelihood is a linear function of the sufficient statistics given by the numbers of out-edges, in-edges and mutual edges at each vertex. Equivalently, the dyad counts at each vertex are sufficient statistics. At vertex \( v \) there are \( n - 1 \) dyads \((y_{uv}, y_{vu})\) for different vertices \( u \) distinct from \( v \). Let \( D_{vij} \) be the number of these dyads that are equal to \((i,j)\) for \( i = 0,1 \) and \( j = 0,1 \). The four dyad counts are linear combinations of the numbers of out-edges, in-edges and mutual edges according to

\[
D_{v10} = \sum_u y_{vu}(1 - y_{uv}) = y_v - \sum_u y_{uv}y_{vu}
\]

\[
D_{v01} = \sum_u (1 - y_{vu})y_{uv} = y_v - \sum_u y_{uv}y_{vu}
\]

\[
D_{v11} = \sum_u y_{uv}y_{vu}
\]

\[
D_{v00} = \sum_u (1 - y_{vu})(1 - y_{uv}) = n - 1 - y_v - y_v + \sum_u y_{uv}y_{vu}
\]

and one of the four dyad counts is redundant.

A classical social network model introduced by Holland and Leinhardt (1981) is a special case of a dyad independence model obtained by assuming that there is a common
mutuality effect at all the vertices. The resulting model has 2n degrees of freedom and sufficient statistics given by the out-degree and the in-degree of each vertex and the total number of mutual edges in the digraph.

The assumption of independent dyads can be combined with an assumption of partial homogeneity among the vertices. This leads to so-called block models in which the sufficient statistics are given by dyad counts within and between the specified blocks of homogeneous vertices. Homogeneity is generally specified as equal values on some vertex variable.

The assumption of independent dyads can be extended to an assumption of Markov dependence. This means that dyads with no common vertex are independent, but other dyads might be dependent. Such Markov graph models are of exponential type, and their log-likelihood functions are linear in sufficient statistics given by vertex degrees and triad counts. See Frank and Strauss (1986) for details and Robins and Pattison (2005) for extensions.

More general models of exponential type have been extensively used for social networks. Such models are specified by assuming that some network statistics that are supposed to be of special importance occur linearly in the log-likelihood function. This allows for great flexibility in choosing statistics but contributes less towards an understanding of the mechanisms creating the network.

Various attempts to investigate the World Wide Web by network methods have created an interest in huge dynamically changing networks with specified degree distributions. Uniform random graphs and Bernoulli graphs with a common edge probability have degree distributions that are approximately Poisson and very different from the degree distributions found in various large-scale empirical networks on the internet and elsewhere. Network models with specified degree distributions can be obtained in various ways, and different approaches are suggested in current research. See Bonato (2008) and Durrett (2007) for references and results on degree distributions and graph processes.

Network data with categorical vertex- and edge-variables can be represented as colored graphs and modeled by block models of various kinds. If there are numerical vertex- and edge-variables, statistical regression models might seem desirable. An example with multivariate normally distributed vertex- and edge-variables can illustrate a modeling approach. The following example is given in Frank (2005). Assume that there is a bivariate vertex-variable \( X = (X_1, X_2) \) where the two components represent out- and in-effects like activity and attraction between individuals, for instance. There is also an edge-variable \( Y \) measuring some strength or duration of specified relations. Different vertices have independent identically distributed vertex-variables \( X(v) = (X_1(v), X_2(v)) \) for \( v = 1, \ldots, n \), and the common distribution is the bivariate normal \( N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \). Conditionally on the vertex-variables and for 1 \( \leq u < v \leq n \), the dyads \( (Y(u,v), Y(v,u)) \) are independent with a bivariate normal distribution that depends on no other vertex-variables than \( X(u) \) and \( X(v) \). This dependence is given by linear regressions on the components of \( X(u) \) and \( X(v) \). The assumptions imply a model of exponential type with easily interpretable parameters. Few attempts seem to have been made to explore the potential of such models in specific applications, and various regression models for networks deserve to be further studied.
4. Network Sampling

Consider a fixed network represented by a vertex-variable

\[ x = \{ (v, x_v) : v \in V \} \]

and an edge-variable

\[ y = \{ (u, v, y_{uv}) : (u, v) \in V^2 \} \].

The vertices are units in a large finite population of labeled units, say \( V = \{1, \ldots, N\} \). Often a preferable alternative to making a census is to use a sample survey and make partial observations of network data. A random subnetwork is observed by selecting a subset \( S \) of \( V \) by a random sampling design and observing the restriction of \( x \) to \( S \) and a restriction of \( y \) to some subset of \( V^2 \) generated by \( S \). The subnetwork within \( S \) defined by

\[ x(S) = \{ (v, x_v) : v \in S \} \]

and

\[ y(S) = \{ (u, v, y_{uv}) : (u, v) \in S^2 \} \]

is often said to be induced by \( S \). Another subnetwork generated by \( S \) is the out-star from \( S \) defined by

\[ x(S) = \{ (v, x_v) : v \in S \} \]

and

\[ y(S) = \{ (u, v, y_{uv}) : u \in S, v \in V \} \]

The in-star to \( S \) is analogously defined, and the star generated by \( S \) is the union of the out-star and the in-star. Also other subnetworks might be observed when vertices or other network elements are sampled.

The random sampling of vertices can be based on a sequential procedure selecting a sample sequence \( (v_1, \ldots, v_n) \) of \( n \) vertices that might contain repeated selections of the same vertex. For sequential sampling procedures, \( S \) is defined as the set of distinct vertices in the sample sequence. The probabilistic sampling design of \( S \) is specified by its inclusion probabilities

\[ \pi(A) = P(A \text{ included in } S) \]

for subsets \( A \) of \( V \). For homogeneous sampling designs, the inclusion probabilities are equal for all subsets of the same size, and we denote by \( \pi_k \) the inclusion probability of any particular subset of size \( k \) for \( k = 1, 2, \ldots \). For simple random sampling of \( n \) vertices selected with or without replacement and for Bernoulli sampling with a common selection probability \( p \), the inclusion probabilities are given by

\[ \pi_k = \binom{n}{k}/\binom{N}{k} \] for \( k = 1, \ldots, n \) (without replacement)

\[ \pi_k = \sum_{j=0}^{\min(k,n)} (-1)^j \binom{k}{j} \binom{n-j}{n-k} \] for \( k = 1, \ldots, n \) (with replacement)

\[ \pi_k = p^k \] for \( k = 1, \ldots, N \) (Bernoulli).
Other sampling designs used in network sampling as well as in ordinary survey sampling might depend on known data for the population units. In network sampling it is also possible to use sampling designs that depend on properties of the network that are not known or even observable in sample data. For instance, so called respondent-driven sampling that has been used for investigating hidden populations, such as drug users or homeless people, for which no sampling frames are available is a design that is based on unknown acquaintanceships in the population. When the sampling design depends on unknown population properties, the statistical methods used are not so distinctly separated into design-based or model-based methods as is common in ordinary survey sampling. The possibilities using sampling designs that are not totally controlled by the investigator and that contain unknown quantities to be estimated by sample data can provide unexpected progress in special situations.

The following is an example referring to official crime statistics, which is discussed in Frank and Carrington (2007). Statistics on co-offending and re-offending are only available for identified offences and identified offenders. Many offences are not reported or identified. Identified offences are reported as cleared if at least one offender is identified, but there might be unidentified co-offenders. Identified offenders might have committed both identified and unidentified reoffences. True distributions of offences according to co-offender size and true distributions of offenders according to reoffending activity are therefore poorly estimated by using available statistics ignoring uncertainty caused by unidentified offences and offenders. A model approach based on network sampling assumes that for each offence, the selection of co-offenders is obtained as a general Bernoulli sample with selection probabilities proportional to reoffending activities. This assumption together with some further assumptions about offence identification and co-offender identification imply that the true distributions of co-offender size and reoffending activity can be estimated, and the results compare favorably with facts known from victimization surveys and other sources.

Even if there is a floating borderline between design and model in network surveys due to the possible use of model assumptions in some sampling designs, it is still meaningful to maintain the distinction between the roles that models can play. As in ordinary survey sampling, uncertainty caused by other factors than sampling can be modeled. Measurement errors and response uncertainty are such factors that are usually modeled. In network surveys, the model approach assumes that the edge-variable $y$ in the population network is the outcome of a random network $Y$ on the same $V^2$. The random network consisting of the restriction of $Y$ to the observable subset of $V^2$ generated by a vertex sample $S$ is random due to both the randomness of the model for $Y$ and the randomness of the probabilistic sampling design for $S$. The design as well as the model might contain unknown parameters that might be estimated by sample data. If they cannot be estimated they might still be useful for a sensitivity analysis in which they are varied in order to investigate their influence on results obtained. The model for $Y$ can be referred to as a super-population network in analogy with the use of the term super-population in ordinary survey sampling.

In ordinary survey sampling it is common to estimate a population total

$$T = \sum_{v \in V} y_v$$
by an unbiased Horvitz-Thompson estimator
\[ T' = \sum_{v \in S} y_v / \pi(v) \]
based on a sample \( S \) obtained according to a design with single-unit inclusion probabilities \( \pi(v) \) for units \( v \in V \). When network data are observed for the subnetwork induced by a vertex sample \( S \), the edge-variable total
\[ T = \sum_{u \in V} \sum_{v \in V} y_{uv} \]
can be analogously estimated by an unbiased Horvitz-Thompson estimator
\[ T' = \sum_{u \in S} \sum_{v \in S} y_{uv} / \pi(u,v) \]
where \( \pi(u,v) \) is the inclusion probability of \( \{u,v\} \). In particular, if \( S \) has a homogeneous design,
\[ T' = \sum_{v \in S} y_{vv} / \pi_1 + \sum_{u \in S} \sum_{v \in S: u \neq v} y_{uv} / \pi_2 \]
The variance of the estimators and unbiased variance estimators can also be obtained using inclusion probabilities of higher orders. For homogeneous designs, all formulae simplify a lot, and inference can be based on a few summary statistics for the observed subnetwork.

When the sampling designs are based on network properties, it can be difficult to determine exact inclusion probabilities required for network inference. A sampling design called snowball sampling starts with an initial vertex sample \( S_0 \) obtained according to some specified design and proceeds by successively asking sampled vertices (people or units that can be represented by spokespersons) for further recruitments. Let \( A(S_0) \) denote a sampled subset of vertices in \( V \) that are adjacent to at least one of the vertices in \( S_0 \) according to a specified graph. The vertices in \( A(S_0) \) that are not in \( S_0 \) constitute the first wave \( W_1 \) of the snowball sample. The vertices in \( A(W_1) \) that are not in \( S_0 \) nor in \( W_1 \) constitute the second wave \( W_2 \), and so forth. A snowball sample with \( k \) waves is given by
\[ S_k = S_{k-1} \cup W_k \]
for \( k = 1, 2, \ldots \). It is usually possible to specify the inclusion probabilities for \( S_0 \) and \( S_1 \) but not for snowball samples with several waves, and it might be convenient to use an approximation. Frank (2010) gives the following approximations to first and second order inclusion probabilities for the successive snowballs when they are selected according to general (nonhomogeneous) Bernoulli sampling both for the initial sample and for the subsequent recruitment of waves. Assume that \( S_0 \) is a general Bernoulli sample with selection probabilities \( p_v \) for \( v \in V \). Assume that the successive waves \( W_k \) are obtained as general Bernoulli samples recruited from vertices \( u \) in the previous wave \( W_{k-1} \) with selection probabilities \( p_{uv} \) for \( v \in V, k = 1, 2, \ldots \) and \( W_0 = S_0 \). It is convenient to define \( p_{vv} = 1 \) for \( v \in V \). Then the successive inclusion probabilities of the first order are given by
\[ \pi_{0,v} = P(v \in S_0) = p_v \]
for the initial sample, and by
\[ p_{1,v} = P(v \in S_1) = 1 - \prod_{u \in V} (1 - p_u \ p_{uv}) \]
for the one-wave snowball. In subsequent \( k \)-wave snowballs the first order inclusion probabilities are recursively given by the approximations
\[ p_{k+1,v} = P(v \in S_{k+1}) = p_{k,v} + \sum_{u \in V} (p_{k,u} - p_{k-1,u}) p_{uv} \]
for \( k = 1, 2, \ldots \) and \( v \in V \). The inclusion probabilities of the second order for the initial sample and for the one-wave snowball are given by
\[ p_{0,v,w} = P(v \in S_0, w \in S_0) = p_v \ p_w \]
\[ p_{1,v,w} = P(v \in S_1, w \in S_1) = p_{1,v} + p_{1,w} - 1 + \prod_{u \in V} [1 - p_u (1 - q_{uv} \ q_{uw})] \]
and in subsequent snowballs they are recursively given by the approximations
\[ p_{k+1,v,w} = p_{k+1,v} - p_{k,v} + p_{k+1,w} - p_{k,w} + p_{k,v,w} - \sum_{u \in V} [(p_{k,u} - p_{k-1,u}) \times (1 - q_{uv} q_{uw})] \]
where \( q_{uv} = 1 - p_{uv} \) for \( k = 1, 2, \ldots \) and \( v \neq w, v \in V, w \in V \).

5. References
