SURVEY SAMPLING IN GRAPHS

Cye FRANK
University of Lund, Lund, Sweden

Received 15 November 1976; revised manuscript received 2 May 1977
Recommended by T. Daleltenus

The Horvitz-Thompson estimation theory is applied to snowball sampling and some other sampling procedures using a known or unknown graph structure in the survey population. In particular, simple graph-parameter estimators and variance estimators are obtained which are based on various kinds of partial information about the graph.

Key words:
Population structure, Graph, Sociogram, Network, Snowball sampling, Estimation

1. Introduction

Graph theory is a branch of finite mathematics that has been powerfully expanded due to the increased use of computers. Graph theory provides concepts and tools that are particularly convenient for analyzing the relationships and interferences between the components in large scale systems, and various graph models have been developed in the many disparate fields of applied mathematics. Noitemeier (1976) is a recent textbook on graph theory which contains a comprehensive list of references and many illustrative applications.

The use of graph models implies a demand for methods for evaluating the models and estimating and testing various graph parameters. The statistical analysis of data in graph models offers a rich variety of problems for research and development, and it is still in its infancy. Some different lines of progress can be distinguished.

Probability problems concerning various properties of a stochastic graph have been considered in a number of early papers, e.g. Erdős, Rényi (1959, 1960), Erdős (1959, 1961), Rényi (1959), Gilbert (1959, 1961) and Austin, Fagon, Penney, Riordan (1959). In later years quite a large number of papers on stochastic graphs have appeared, in particular in Theory of Probability and Its Applications. A few further references are Frank (1968), Stepanov (1969), Ivchenko (1973), Ling (1973, 1975), Robinson, Schwenk (1975) and Tainiter (1975).

Another approach which is oriented towards the numerical algorithms is provided by the research on clustering methods, multidimensional scaling methods and other explorative statistical methods of search for structure in empirical data. Matula (1970) and Hubert (1974) describe the applications of graph theory to cluster analysis. The stochastic models have not yet received any advanced position in these kinds of data analysis. An outstanding reference is Ling (1973) who suggests a stochastic approach to cluster analysis.
The field of survey sampling and statistical inference in a graph has been considered by only a few authors so far. Snowball sampling has been treated by Goodman (1961), Bloemena (1964) has derived moments of the sampling distributions of some graph statistics, and Proctor (1966, 1967) has estimated the edge frequency of a graph. Some generalizations and extensions have been given by Sheridan (1970) and Proctor, Suwattee (1970). An interesting paper by Stephan (1969) discusses the possible future development of survey sampling and the need to utilize the known or observable structure that may be inherent in a population. Various graph inference problems have been considered by Frank (1969a, b, 1971, 1975a, b, 1976, 1977), Capobianco (1970, 1972, 1974), Tapiero, Boots (1974) and Tapiero, Capobianco, Lewin (1975). There are some papers treating various sampling and inference problems which are not explicitly stated as graph problems but which can be given such formulations. Some relevant papers are dealing with matching, overlapping and multiplicity. See for instance Goodman (1952), Deming and Glasser (1959), Ghosh (1966), Frank (1970), Sirken (1970, 1972a, b, c), Sirken and Levy (1974) and Nathan (1976).

In this paper we shall consider the Horvitz–Thompson estimation theory in survey sampling and show how this theory can be applied in graph models. Some basic graph concepts are given in Section 2, and some preliminaries on Horvitz–Thompson estimation are given in Section 3. Section 4 discusses the use of snowball sampling in surveys, and Section 5 gives some estimators of a population total based on snowball sample data. The questions of observability and data accessibility in the graph which generates the snowball sample are discussed in Section 6. This discussion leads to the introduction of a graph total which is a generalization of the population total. Section 7 shows how the Horvitz–Thompson theory can be applied to graph totals. Section 8 gives a number of examples illustrating the use of graph totals for describing various graph properties. Various sampling and observation procedures are also illustrated. Finally Sections 9–12 are devoted to the estimation of graph totals based on various kinds of sample data.

2. Some basic graph concepts

A directed (undirected) graph consists of a non-empty set $V$ of elements called vertices (nodes, points), a set $E$ of elements called edges (arcs, links, lines, arrows) and an incidence function $I$ which associates an ordered (unordered) pair of vertices $!e)$ with each edge $e \in E$. Let $E_{ij}$ denote the subset of edges which are incident with the ordered (unordered) vertex pair $(i, j) \in V^2$. It is convenient to use the notation $E_{ij} = E_{ji}$ in the undirected case, and permitting a slight abuse of notations we can put

$$I(e) = \{(i, j) \in V^2 : e \in E_{ij}\}. \quad (1)$$

In the undirected case the unordered pair $\{i, j\}$ will thus be identified with the pair of ordered pairs $\{(i, j), (j, i)\}$. The sets $E_{ij}$ may be empty for some or all $(i, j) \in V^2$. For a directed graph the non-empty $E_{ij}$ constitute a partition of $E$, i.e. they are disjoint and exhaustive of $E$. For an undirected graph $E_i = E_{ji}$ for all $(i, j) \in V^2$, and the non-empty distinct $E_{ij}$ constitute a partition of $E$.

The edges in $E_{ij} \cup E_{ji}$ are said to be between $i$ and $j$, and $i$ and $j$ are called their
endvertices. The edges in $E_{ij}$ are also said to lead from $i$ to $j$, and $i$ and $j$ are called their initial and terminal vertices, respectively. If there are two or more edges in $E_{ij}$, they are said to be parallel. The edges in $E_i$ are called the loops at $i$. The edges in

$$E_i = \bigcup_{j \in V} E_{ij} \quad \text{and} \quad E_{-i} = \bigcup_{j \in V} E_{ji}$$

(2)

are called the outedges and the inedges at $i$, respectively. In the undirected case the outedges and the inedges coincide.

It is convenient to use the incidence concept not only for edges and vertex pairs, but also for edges and vertices and for edges with a common endvertex. Thus two vertices $i$ and $j$ are said to be incident if there is an edge between them. In the directed case $i$ is also said to be incident before $j$, and $j$ is said to be incident after $i$, if there is an edge from $i$ to $j$. The incident edges at a vertex $i$ comprise the loops, the outedges and the inedges at $i$. The incident edges of an edge $e \in E_{ij}$ are the incident edges at $i$ and the incident edges at $j$. The incident vertices of an edge are its endvertices. Let $I_1(e)$ and $I_2(e)$ denote the sets of the initial and terminal vertices of $e$, respectively, i.e.

$$I_1(e) = \{i \in V : e \in E_{-i}\} \quad \text{and} \quad I_2(e) = \{i \in V : e \in E_i\}.$$  

(3)

For an undirected graph both these sets coincide. For arbitrary edge subsets $M \subseteq E$ we define the sets of initial and terminal incident vertices

$$I_1(M) = \bigcup_{e \in M} I_1(e) = \{i \in V : M \cap E_i \neq \emptyset\}$$

(4)

$$I_2(M) = \bigcup_{e \in M} I_2(e) = \{i \in V : M \cap E_{-i} \neq \emptyset\}$$

where $\emptyset$ is the empty set. Analogously we define the set of incident vertex pairs

$$I(M) = \bigcup_{e \in M} I(e) = \{(i, j) \in V^2 : M \cap E_{ij} \neq \emptyset\}.$$  

(5)

Two vertices $i$ and $j$ are said to be adjacent if they are equal or incident. In the directed case $i$ is also said to be adjacent before $j$, and $j$ is said to be adjacent after $i$, if either $i = j$ or there is an edge from $i$ to $j$. The sets of vertices adjacent after and before $i$ are denoted by

$$A_i = \{i\} \cup \{j \in V : E_{ij} \neq \emptyset\} \quad \text{and} \quad B_i = \{i\} \cup \{j \in V : E_{ji} \neq \emptyset\},$$

(6)

respectively. The set of the adjacent vertices at $i$ is $A_i \cup B_i$. The set of vertices adjacent after at least one of the vertices in a subset $M \subseteq V$ and the set of vertices adjacent before at least one of the vertices in a subset $M \subseteq V$ are denoted by

$$A(M) = \bigcup_{i \in M} A_i \quad \text{and} \quad B(M) = \bigcup_{i \in M} B_i,$$

(7)

respectively.
A sequence \((e_1, \ldots, e_n) \in E^n\) of \(n\) edges is called a **walk** of **length** \(n\) if there exists a sequence \((i_0, \ldots, i_n) \in V^{n+1}\) of \(n+1\) vertices such that

\[
e_k \in E_{i_{k-1}i_k} \quad \text{for} \quad k = 1, \ldots, n.
\] (8)

The edges \(e_1, \ldots, e_n\) and the vertices \(i_0, \ldots, i_n\) are said to be **incident** with the walk. The walk is said to be **between** \(i_0\) and \(i_n\) and **through** \(i_1, \ldots, i_{n-1}\). The walk is also said to lead from \(i_0\) to \(i_n\) and \(i_0\) and \(i_n\) are called the **endvertices** of the walk. If \(i_0 = i_n\) the walk is said to be **closed**. A walk in which all the incident edges are different is said to be **simple**. A walk in which all the incident vertices are different, except for possible equality between the endvertices, is called a **path**. Each path is a simple walk. If a path is a closed walk it is called a **closed path**. A simple closed walk need not be a closed path. If there is a walk of length \(n\) from \(i\) to \(j\), \(j\) is said to be **reachable** in \(n\) steps from \(i\).

Two vertices \(i\) and \(j\) are said to be **connected** if either \(i = j\) or if there is a walk from \(i\) to \(j\) and a walk from \(j\) to \(i\). This concept of connectedness defines an equivalence relation in \(V\), and its equivalence classes are called the **connected components** of \(V\). Any two different vertices in a connected component are incident with a closed walk (and with a path but not necessarily with a closed path). No two vertices in different connected components are incident with a closed walk. Consequently the edges which are not incident with any closed walk are exactly the edges which are between vertices in two different connected components.

The connected components in a directed graph are called **strongly connected components** by some authors. **Weakly connected components** are then defined as the connected components which are obtained if the directed graph is considered as an undirected graph, i.e. if the orders in the incident vertex pairs of the edges are ignored.

A graph is said to be **simple** if it has no loop and at most one edge incident with each vertex pair. In a simple directed graph two edges are said to be **mutual** if they are incident with the same vertices.

A graph specified by the vertex set \(V\) and the sets \(E_{ij}\) of edges incident with \((i,j) \in V^2\) has a **subgraph** specified by the vertex set \(V'\) and the sets \(E'_{ij}\) of edges incident with \((i,j) \in (V')^2\) if

\[
\begin{align*}
V' & \subseteq V \\
E'_{ij} & \subseteq E_{ij} \quad \text{for} \quad (i,j) \in (V')^2 \\
E_{ij} & = E_{ji} \quad \text{implies} \quad E'_{ij} = E'_{ji} \quad \text{for} \quad (i,j) \in (V')^2.
\end{align*}
\] (9)

The **subgraph generated by a vertex subset** \(M \subseteq V\) is defined as the subgraph with

\[
V' = M \quad \text{and} \quad E'_{ij} = E_{ij} \quad \text{for} \quad (i,j) \in (V')^2.
\] (10)

The **subgraph generated by an edge subset** \(M \subseteq E\) is defined as the subgraph with

\[
V' = I_1(M) \cup I_2(M) \quad \text{and} \quad E'_{ij} = M \cap E_{ij} \quad \text{for} \quad (i,j) \in (V')^2.
\] (11)
In particular, if \( M \subseteq E \) then the subgraph generated by the vertex subset \( I_1(M) \cup I_2(M) \) is equal to the subgraph with
\[ V' = I_1(M) \cup I_2(M) \quad \text{and} \quad E'_{ij} = E_{ij} \quad \text{for} \quad (i, j) \in (V')^2. \tag{12} \]
The two subgraphs given by (11) and (12) are identical if the graph is simple.

3. Preliminaries concerning survey sampling

We shall consider a finite population of \( N \) units and a real variable \( y \) defined in the population. Let \( V = \{1, \ldots, N\} \) denote the population, let \( y_i \) denote the value of \( y \) for unit \( i \in V \) and let the total of the \( y \)-values be denoted by \( T = \sum y_i \).

A stochastic variable \( S \) taking values in the set of subsets of \( V \) is called a random sample, and the probability distribution giving non-negative selection probabilities \( P(S = M) = p(M) \) to each subset \( M \) of \( V \) is called the sampling design. The random sample \( S \) can be represented by a vector \((S_1, \ldots, S_N)\) of \( N \) stochastic variables defined by
\[ S_i = \begin{cases} 1 & \text{if } i \in S, \\ 0 & \text{otherwise} \end{cases} \tag{13} \]
for \( i \in V \). The stochastic variable \( S_i \) indicates whether unit \( i \) is included in the random sample or not, and it is called an inclusion indicator of \( S \). Obviously the selection probabilities can be obtained as the expected value
\[ p(M) = \mathbb{E} \prod_{i \in M} S_i \prod_{j \in \bar{M}} (1 - S_j), \tag{14} \]
where \( \bar{M} \) denotes the complement of \( M \). The sampling design which is specified by the selection probabilities can alternatively be specified by the inclusion probabilities
\[ \pi(M) = \mathbb{E} \prod_{i \in M} S_i = P(M \subseteq S) \tag{15} \]
or the exclusion probabilities
\[ \bar{\pi}(M) = \mathbb{E} \prod_{i \in M} (1 - S_i) = P(M \subseteq \bar{S}) \tag{16} \]
for all non-empty \( M \subseteq V \). For convenience we define \( \pi(\emptyset) = \bar{\pi}(\emptyset) = 1 \) where \( \emptyset \) is the empty set. The inclusion and exclusion probabilities are also denoted by \( \pi_i, \bar{\pi}_i, \ldots \) and \( \bar{\pi}_{ij}, \bar{\pi}_{ij}, \ldots \) if \( M = \{i\}, \bar{M} = \{i, j\}, \ldots \)

Proving the identity
\[ \prod_{i \in M} (1 - S_i) = \sum_{L \subseteq M} (-1)^{|L|} \prod_{i \in L} S_i \tag{17} \]
is a straightforward matter. Here \(|L|\) denotes the number of elements in \( L \), the sum is over all subsets of \( M \), and the term corresponding to \( L = \emptyset \) is understood to be 1. Taking expectations in (17) entails that the exclusion probability \( \bar{\pi}(M) \) can be obtained from
the inclusion probabilities according to the formula

\[ \pi(M) = \sum_{L \subseteq M} (-1)^{|L|} \pi(L). \] (18)

In particular

\[ \pi_i = 1 - \pi_i \]
\[ \pi_{ij} = 1 - \pi_i - \pi_j + \pi_{ij}. \] (19)

Analogously we can express \( \pi(M) \) in terms of the exclusion probabilities:

\[ \pi(M) = \sum_{L \subseteq M} (-1)^{|L|} \pi(L). \] (20)

The selection probabilities \( p(M) \) can be obtained from the inclusion or the exclusion probabilities according to

\[ p(M) = \sum_{L \subseteq M} (-1)^{|L|} \pi(L \cup M) = \sum_{L \subseteq M} (-1)^{|L|} \pi(L \cup M). \] (21)

We also note the following monotonic properties of \( \pi \) and \( \pi \):

\[ \pi(L) \geq \pi(M) \quad \text{and} \quad \pi(L) \geq \pi(M) \quad \text{if} \quad L \subseteq M. \] (22)

If the inclusion indicators \( S_1, \ldots, S_N \) of the random sample \( S \) are independent stochastic variables with \( E S_i = p_i \) for \( i \in V \), then the random sample is said to have a Bernoulli \( (p_1, \ldots, p_N) \) design. If \( p_i = p \) for \( i \in V \) the design is called a simple Bernoulli \( (p) \) design. We will use this design to illustrate and simplify the general results in this paper. It should be noticed that for practical purposes it is often satisfactory to approximate a simple random sampling design with a simple Bernoulli \( (p) \) design where \( p \) is the sampling fraction.

For later use we shall also give some well-known results concerning Horvitz-Thompson estimation in survey sampling. See for instance Kendall and Stuart (1966).

If \( \pi_i > 0 \) for all \( i \in V \), then the total \( T \) has an unbiased estimator, the Horvitz-Thompson estimator, given by

\[ \hat{T} = \sum_{i \in S} \frac{y_i}{\pi_i}, \] (23)

and the variance of \( \hat{T} \) is given by

\[ \sigma^2(\hat{T}) = \sum_{i \in V} \sum_{j \in V} y_i y_j \pi_{ij}. \] (24)
where

\[ y_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \]  

(25)

and \( \pi_{ii} = \pi_i \) for \( i \in V \). If \( \pi_{ij} > 0 \) for all \( i \in V \) and \( j \in V \), then the variance has an unbiased estimator given by

\[ \hat{\delta}^2(T) = \sum_{i \in S} \sum_{j \in S} y_{ij} y_{j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_i \pi_j}, \] 

(26)

the Horvitz-Thompson variance estimator. Other variance estimators, such as for instance the Yates-Grundy variance estimator, will not be considered in this paper, and no attempt has been made to evaluate or compare various variance estimators.

4. Some illustrations of snowball sampling

Assume that a graph is defined with the population \( V = \{1, \ldots, N\} \) as its vertex set. By use of the graph it is possible to extend an initial random sample \( S \subseteq V \) by joining the vertices which are adjacent after at least one vertex in \( S \). The extended random sample becomes \( S' = A(S) \). Now \( S' \supseteq S \), and if this inclusion is strict, \( S' \supsetneq S \), then a second stage is given by \( S'' = A(S') \supseteq S' \). If \( S'' \supsetneq S' \) a third stage is given by \( S''' = A(S'') \supseteq S'' \), etc. The sequence \( S', S'', \ldots \) of successively extended samples is called a sequence of snowball samples. The following discussion will mostly be restricted to one stage snowball samples \( S' \). We shall indicate some situations where snowball samples might be of interest.

Assume that the population can be partitioned into equivalence classes according to some kind of prior knowledge. A simple illustration is the partition of a population of individuals into households. Formally the household structure can be considered as a graph with the vertices as the individuals and the edges connecting the individuals which are members of a common household. This graph is undirected, has a loop at each vertex, has no parallel edges, and has an edge between any two vertices in the same connected component. The equivalence relation “is a member of the same household as” is an example of a relation which is ordinarily known prior to the selection of the initial sample and consequently usable for this selection procedure.

It is also possible to have an equivalence relation which is unknown but observable in the course of the survey. For instance the relations “belongs to the same opinion group as” or “has the same background as” might be relations of this kind.

If the individuals in a population are supposed to report about the occurring edges in a graph representing an equivalence relation in the population, then it is natural to assume that the knowledge about the occurring edges is not complete and perhaps not free from error. The individuals might not be aware of some of their edges, and the graph might then represent a partially known equivalence relation. The connected components of such a graph are equal to the equivalence classes or are subsets of them.

The prior knowledge concerning a population can also consist of more general relations than equivalence relations. For instance, there may be knowledge about
contacts or dependencies between the units in a population which can be represented by a directed graph. The contacts can be cooperation contacts, friendship, kinship, economic dependencies, etc. In the same way as for the equivalence relations the contacts can be considered to be known and usable for the drawing of the initial sample, or the contacts can be unknown and partially observable in the survey.

In the ordinary sample survey the sample data consist of the observed \( y \)-values for the units in the sample \( S \), i.e., \( y_i \) for \( i \in S \). When considering a snowball sample one has to distinguish between various kinds of information concerning the graph as has been indicated above. The graph can be completely or partially known prior to the survey. It can be possible to make observations of certain graph properties simultaneously as the \( y \)-values are observed for the units in the sample. We shall give two simple examples.

The sample information can be collected by indirect interviews. Then the information about the \( y \)-values of all the individuals in the snowball sample \( S' \) is obtained from interviews with the individuals in the initial sample \( S \). For instance each member of a household might report the \( y \)-value of any member of the household. Sample data then consist of \( y_j \) for \( j \in S' \) and \( x_{ij} \) for \( i \in S \) and \( j \in V \) where

\[
x_{ij} = \begin{cases} 1 & \text{if } i \text{ reports the } y \text{-value of } j, \\ 0 & \text{otherwise}. \end{cases} \tag{27}
\]

Another possibility is that the individuals in \( S \) give information about their \( y \)-values and their adjacencies, and that the information about the \( y \)-values of the other individuals in \( S' \) is then obtained by direct interviews with them. In this case the sample data consist of \( y_j \) for \( j \in S' \) and \( x_{ij} \) for \( i \in S' \) and \( j \in V \) where

\[
x_{ij} = \begin{cases} 1 & \text{if } i \text{ suggests an interview with } j, \\ 0 & \text{otherwise}. \end{cases} \tag{28}
\]

If the snowball sampling is continued with a second stage the sample data will consist of \( y_j \) for \( j \in S'' \) and \( x_{ij} \) for \( i \in S'' \) and \( j \in V \), and so forth if several stages are considered.

5. Estimation of a total from a snowball sample

Assume that the graph which is used in forming the snowball sample \( S' = A(S) \) is completely known. The adjacency indicators are defined by

\[
x_{ij} = \begin{cases} 1 & \text{if } i \text{ is adjacent to } j, \\ 0 & \text{otherwise}. \end{cases} \tag{29}
\]

Let us denote the selection probabilities of the sampling design of \( S' \) by

\[
P(S' = M) = p'(M) \tag{30}
\]
and the inclusion and exclusion probabilities by

\[ P(M \subseteq S') = \pi'(M), \]
\[ P(M \subseteq S) = \bar{\pi}(M). \]  \hfill (31)

The following lemma gives these probabilities in terms of the corresponding probabilities for the initial random sample S.

**Lemma 1.** The sampling design of a snowball sample \( S' = A(S) \) with an initial random sample S and a fixed graph has the selection, inclusion and exclusion probabilities given by

\[ p'(M) = \Sigma_{L \subseteq M} (-1)^{|L|} \bar{\pi}(B(L \cup \overline{M})), \]  \hfill (32)
\[ \pi'(M) = \Sigma_{L \subseteq M} (-1)^{|L|} \bar{\pi}(B(L)), \]  \hfill (33)
\[ \pi'(M) = \bar{\pi}(B(M)) \]  \hfill (34)

for subsets \( M \subseteq V \).

**Proof.** The inclusion indicators of \( S' \) defined by

\[ S'_j = \begin{cases} 1 & \text{if } j \in S' \\ 0 & \text{otherwise} \end{cases} \]  \hfill (35)

can be obtained from the inclusion indicators of S and the adjacency indicators of the graph according to

\[ S'_j = \max_{i \in V} S_{ij} x_{ij} = 1 - \prod_{i \in B_j} (1 - S_i). \]  \hfill (36)

It follows that

\[ \prod_{j \in M} (1 - S'_j) = \prod_{i \in B(M)} (1 - S_i) \]  \hfill (37)

and by taking expectations we obtain (34). Substituting (34) into (20) applied to \( S' \) gives (33), and substituting (34) into (21) applied to \( S' \) gives (32).

The following theorem shows that if we have recourse to snowball sample data then we can estimate the total \( T \) by an unbiased estimator which dominates the Horvitz-Thompson estimator \( \hat{T} \) based on the initial random sample S.
Theorem 2. The Horvitz–Thompson estimator $\hat{T}$ based on the initial random sample $S$ is dominated by the unbiased estimator

$$T^* = \sum_{i \in S'} \frac{y_i}{\pi_i} \left[ \frac{\sum_{L \in S'} (-1)^{|L|} \pi((i) \cup B(L \cup S'))}{\sum_{L \in S'} (-1)^{|L|} \pi(B(L \cup S'))} \right]$$

based on the snowball sample $S' = A(S)$.

Proof. We have to prove that

$$ET^* = T \quad \text{and} \quad \sigma^2(T^*) \leq \sigma^2(\hat{T}) \quad \text{for all } (y_1, \ldots, y_N).$$

We shall show that $T^*$ is equal to the conditional expectation $T^* = \mathbb{E}(\hat{T}|S') = \sum_{i \in S'} \frac{y_i}{\pi_i} \mathbb{E}(S_i|S')$, and then the unbiasedness and the variance inequality will follow from the elementary rules of calculation for conditional stochastic variables. Now $S_i = 0$ implies that $S_i = 0$, and it follows that

$$\mathbb{E}(\hat{T}|S') = \sum_{i \in S'} \frac{y_i}{\pi_i} \mathbb{E}(S_i|S').$$

In order to determine the conditional expectation $\mathbb{E}(S_i|S')$ we note that

$$\mathbb{E}(S_i|S' = M') = 1 - \frac{1}{p'(M)} \mathbb{E}(1 - S_i) \prod_{j \in M} S_j \prod_{k \in M} (1 - S_k)$$

$$= 1 - \frac{1}{p'(M)} \mathbb{E} \sum_{L \subseteq M} (-1)^{|L|}(1 - S_i) \prod_{j \in L \setminus M} (1 - S_j)$$

$$= 1 - \frac{1}{p'(M)} \sum_{L \subseteq M} (-1)^{|L|} \pi((i) \cup B(L \cup M))$$

where the last equality follows by application of (37). Using Lemma 1 and substituting $M = S'$ we then obtain the desired result, and the proof is demonstrated.

The estimator $T^*$ depends in a complicated way on the sampling design of the initial sample, and it seems to be desirable to provide some simpler estimator. The following theorem gives the Horvitz–Thompson estimator based on $S'$. 

Theorem 3. If \( \bar{\pi}(B_i) < 1 \) for \( i \in V \) the Horvitz–Thompson estimator of \( T \) based on \( S' \) is given by

\[
\hat{T}' = \sum_{i \in S'} \frac{y_i}{1 - \bar{\pi}(B_i)},
\]

and it has a variance given by

\[
\sigma^2(\hat{T}') = \sum_{i \in V} \sum_{j \in V} y_i y_j \gamma_{ij}.
\]

where

\[
\gamma_{ij} = \frac{\bar{\pi}(B_i \cup B_j) - \bar{\pi}(B_i)\bar{\pi}(B_j)}{[1 - \bar{\pi}(B_i)][1 - \bar{\pi}(B_j)]}.
\]

Proof. Applying (23)–(25) with \( S' \) and using Lemma 1 to obtain \( \pi_i \) and \( \pi_{ij} \) will yield the result.

The following theorem implies that generally one cannot be sure that \( \hat{T}' \) dominates \( \hat{T} \).

Theorem 4. If there exists a pair of distinct vertices \( i \) and \( j \) such that the graph and the initial sampling design satisfy

\[
B_i = \{i\} \subset B_j \quad \text{and} \quad \bar{\pi}(B_j)\pi_{ij} \neq \pi_{ij} - \pi_i\pi_j,
\]

then neither \( \hat{T}' \) nor \( \hat{T} \) dominates the other.

Proof. We have to show that the condition (46) for some \( i \neq j \) implies that the quadratic form

\[
\sigma^2(T) - \sigma^2(\hat{T}') = \sum_{i \in V} \sum_{j \in V} y_i y_j (\gamma_{ij} - \gamma_{ij}')
\]

is not non-negative or non-positive definite. According to well-known facts about quadratic forms it is sufficient to prove that

\[
(\gamma_{ii} - \gamma_{ii}')(\gamma_{jj} - \gamma_{jj}') - (\gamma_{ij} - \gamma_{ij})^2 < 0
\]

for some \( i \neq j \). Now using (25) and (45) implies that

\[
\gamma_{ii} - \gamma_{ii}' = 0 \quad \text{if} \quad B_i = \{i\}
\]
and

\[ \gamma_{ij} - \gamma'_{ij} \neq 0 \quad \text{if (46) is valid.} \quad (50) \]

Consequently (48) is valid and the theorem is proved.

*Note.* To prove the existence of a sampling design satisfying (46) it is sufficient to note that

\[ \hat{\pi}(B_j) = p_i p_j \prod_{k \in B_j} (1 - \gamma_k) \neq 0 \quad (51) \]

if the initial sample has a Bernoulli \((p_1, \ldots, p_N)\) design with \(0 < p_i < 1\) for \(i \in V\).

Theorem 4 shows that generally \(\hat{T}'\) is not dominating \(\hat{T}\) if the graph has a vertex which has no inedge but at least one outedge. The author has not been able to find a necessary and sufficient condition for \(\hat{T}'\) to be dominating \(\hat{T}\).

6. Comments on graph observability

We have hitherto considered the vertex set \(V\) as a population from which we have drawn the initial sample \(S\). We have extended \(S\) to a snowball sample \(S'\) by using a known graph, and we have observed the values of a variable \(y\) for the vertices in the sample \(S'\).

If the graph is unknown but locally observable at the sampled vertices, it may be asked what kind of graph observations are required in order to apply the estimators of the preceding section. An examination of the Horvitz-Thompson estimator \(\hat{T}'\) in Theorem 3 shows that it requires the observability of \(B_i\) for \(i \in S'\). If the initial sampling design is a simple Bernoulli \((p)\) design with \(p = 1 - q > 0\), then

\[ \hat{T}' = \sum_{i \in S'} \frac{y}{1 - q^{x_i}}, \quad (52) \]

and it is sufficient to observe the size \(x_i\) of \(B_i\) for \(i \in S'\).

From Theorem 3 and the application of (26) to \(\hat{T}'\), it follows that the Horvitz-Thompson estimator of \(\sigma^2(\hat{T}')\) also requires the observability of \(B_i\) for \(i \in S'\). If the initial sampling design is a simple Bernoulli \((p)\) design with \(p = 1 - q > 0\) then

\[ \hat{\pi}(B_i \cup B_j) = q^{x_i + x_j - \sum_{k \in \sigma} x_{ki} x_{kj}} \quad (53) \]

and it follows that \(\hat{\sigma}^2(\hat{T}')\) requires the observability of the sizes

\[ x_{ij} \quad \text{and} \quad \sum_{k \in \sigma} x_{ki} x_{kj} \quad (54) \]

of \(B_i\) and \(B_i \cap B_j\) for \(i \in S'\) and \(j \in S'\).
The observability conditions considered above require sample data concerning certain local properties of the entire graph. In many situations, however, it is not possible to observe more than some local properties of a sampled subgraph. For instance \( B_i \) might not be observable but only \( S \cap B_i \). We are now going to study some estimation problems based on partial graph knowledge. We shall introduce a vertex-pair variable \( y \) with values \( y_{ij} \) for \((i, j) \in V^2\). We shall show how various graph characteristics can be described by the totals

\[
T = \sum_{(i,j) \in V^2} y_{ij},
\]

which we will call \textit{graph totals}, and we shall show how \( T \) can be estimated from various kinds of sample data.

7. Estimation of graph totals

We shall consider a graph with the vertex set \( V = \{1, \ldots, N\} \), the edge set \( E \) and the incidence indicators \( x_{ij} \) defined by

\[
x_{ij} = \begin{cases} 
1 & \text{if } i \text{ is incident to } j, \\
0 & \text{otherwise}
\end{cases}
\]

for \((i, j) \in V^2\). Note that the \( x \)'s now refer to incidences and not adjacencies, i.e. the diagonal entries need not be 1. We have a real variable \( y \) with unknown values \( y_{ij} \) for \((i, j) \in V^2\), and we are interested in estimating the graph total \( T \). At our disposal we have the values of \( y \) observed from a sample \( S' \) of vertex pairs, i.e. \( y_{ij} \) for \((i, j) \in S' \) where \( S' \subseteq V^2 \).

The sample \( S' \) can be obtained in various ways from an initial sample \( S \) by using the graph. The initial sample will be selected according to a known sampling design. We shall consider two different kinds of initial sampling units. In Section 9–11 we shall consider the case of an initial vertex sample \( S \subseteq V \) and some alternative ways of defining the final vertex pair sample \( S' \subseteq V^2 \). After that we shall consider in Section 12 an initial edge sample \( S \subseteq E \) and some ways of generating a vertex pair sample \( S' \subseteq V^2 \) from it.

The sampling design of the vertex pair sample \( S' \) depends on the graph and the sampling design of \( S \). It should be noticed that \( \pi_{ij} \) and \( \pi_{ij} \) denote the inclusion and exclusion probabilities of a single vertex pair \((i, j)\), and \( \pi_{ijkl} \) and \( \bar{\pi}_{ijkl} \) denote the inclusion and exclusion probabilities of a pair of vertex pairs \((i, j)\) and \((k, i)\). Thus (19) does not apply without modification. For convenience we shall put \( \pi_{ijkl} = \pi_{ij} \) and \( \bar{\pi}_{ijkl} = \bar{\pi}_{ij} \).

If the inclusion probabilities \( \pi_{ij} \) are positive for \((i, j) \in V^2\) and can be determined from the sample data for \((i, j) \in S'\), then it is possible to apply the Horvitz–Thompson estimation theory to obtain an unbiased estimator of \( T \) as

\[
\hat{T} = \sum_{(i,j) \in S'} \frac{y_{ij}}{\pi_{ij}}.
\]
The variance of this estimator is

\[ \sigma^2(\hat{T}) = \sum_{(i,j) \in V^2} \sum_{(k,l) \in V^2} y_{ij}y_{kl} \gamma_{ijkl} \]  

(58)

where

\[ \gamma_{ijkl} = \frac{\pi_{ijkl} - \pi_{ij}' \pi_{kl}'}{\pi_{ij}' \pi_{kl}'} \]  

(59)

If \( \pi_{ij}' \) is zero for some \((i,j)\), and if \( y_{ij} \) is zero for these \((i,j)\), then the formulae can be applied with the convention that the total is over all pairs for which \( \pi_{ij}' > 0 \).

If the inclusion probabilities \( \pi_{ijkl}' \) are positive for \((i,j) \in V^2 \) and \((k,l) \in V^2 \) and can be determined from the sample data for \((i,j) \in S' \) and \((k,l) \in S' \), then the variance has the unbiased Horvitz-Thompson estimator

\[ \sigma^2(\hat{T}) = \sum_{(i,j) \in S} \sum_{(k,l) \in S} y_{ij}y_{kl} \frac{\gamma_{ijkl}'}{\pi_{ijkl}'} \]  

(60)

These formulae will be applied in the following in order to derive explicit results pertaining to various specifications of the way \( S' \) is generated from \( S \).

8. Some illustrative examples

We shall indicate some concrete situations where the graph totals may be of interest and where the sampling schemes fitting our theory may be applicable.

A contact network or a sociogram is a graph where the vertices represent individuals, households, enterprises or some other type of units which are connected with each other by some kind of contact represented by the edges. If the vertex-pair variable \( y \) is defined as the number of edges which are incident with the vertex pair, i.e. \( y_{ij} = |E_{ij}| \), then the graph total becomes equal to the total number of contacts if these are directed, and twice the total number of contacts if these are undirected and no loops occur.

With the alternative definition \( y_{ij} = x_{ij} \), where \( x_{ij} \) is the incidence indicator, the graph total becomes equal to the number of incident vertex pairs, i.e. the number of ordered pairs of units which have a contact. In particular, \( y_{ij} \) is equal to the incidence indicator \( x_{ij} \) in both the cases considered above if the graph is simple.

If we define \( y_{ij} = x_{ij}x_{ji} \) for a simple directed graph then the graph total becomes equal to the total mutual edge frequency, i.e. twice the total number of reciprocal choices in the terminology of sociometry.

The following comprise some further examples of structure parameters that may be of interest in contact networks. In an undirected sociogram the total number of isolated pairs of units with a contact can be obtained as a graph total by defining

\[ y_{ij} = \begin{cases} x_{ij} & \text{if } x_i = x_j = x_{ij}, \\ 0 & \text{otherwise}. \end{cases} \]  

(61)
The total number of isolated units in an undirected sociogram can be obtained as a graph total (which is a vertex-value total) by defining

\[
y_{ij} = \begin{cases} 
1 & \text{if } i = j \text{ and } x_{ij} = 0, \\
0 & \text{otherwise}. 
\end{cases}
\]  

(62)

In a directed sociogram the total number of contacts, which are taken by a unit which receives contacts from \(r\) units, and are taken with a unit which emits contacts to \(s\) units, can be obtained as a graph total by defining

\[
y_{ij} = \begin{cases} 
|E_{ij}| & \text{if } x_{ij} = r \text{ and } x_{ij} = s, \\
0 & \text{otherwise}. 
\end{cases}
\]  

(63)

In particular, in a simple directed graph this graph total becomes equal to the number of edges from a vertex with \(r\) inedges to a vertex with \(s\) outedges.

Let us now consider some different ways of observing a part of a large contact network. Assume that a sample \(S\) of units is selected. Information about the occurrences of contacts for the sampled units is obtained from interviews with them. It may be convenient to ask about the contacts within the sample, for instance by presenting a list of \(S\) and asking the units to indicate their contacts there. This may reduce the risk of missing contacts. Then the sample information comprise \(E_{ij}\) for \((i, j) \in S^2\), i.e. the subgraph generated by the vertex sample \(S\) is observed. Obviously, for some of the \(y\)-variables above, it is not sufficient to observe only the contacts within the sample. By asking the units in the sample \(S\) about their contacts within the whole population \(V\) the information is extended to \(E_{ij}\) for \((i, j) \in S \times V \cup V \times S\).

The observation procedures can also be based on samples of contacts. The contacts can for instance be letters, phone calls or some other objects or occurrences which are possible to record and sample. If \(S\) is a sample of contacts it may be convenient to ask the incident units, the units in \(I(S) \cup I(S)\), about their contacts. The information obtained from \(S\) might then consist of \(E_{ij}\) for

\[
(i, j) \in [I_1(S) \cup I_2(S)] \times V \cup V \times [I_1(S) \cup I_2(S)].
\]

Flow networks are graph models which are used to describe transportation and communication systems in a wide sense. The flows between the vertices can be amounts of money transferred between different accounts, quantities of goods distributed from producers to consumers, information, rumour or infection communicated between individuals, etc. Let \(f_{ij}\) denote the amount of flow transmitted from vertex \(i\) to vertex \(j\) for \(i \neq j\), and let \(g_i\) and \(h_i\) denote the amounts of flow which are generated and absorbed, respectively, at vertex \(i\). The generated and absorbed flows at vertex \(i\) can be the initial and the final amounts of a certain period of time, or they can be the flows to vertex \(i\) from the sources outside the considered vertex set \(V\) and the flows from vertex \(i\) to the receivers outside \(V\), respectively.

In a goods distribution system the total amount of transportation work can be obtained as a graph total if \(y_{ij}\) is defined as the product of the amount of goods \(f_{ij}\) and the
the distance $d_{ij}$ from $i$ to $j$, i.e.,

$$y_{ij} = \begin{cases} f_{ij}d_{ij} & \text{if } i \neq j, \\ 0 & \text{otherwise}. \end{cases}$$  \hspace{1cm} (64)

If re-loading of the incoming and outgoing goods is required at each vertex, then the total amount of re-loaded goods can be obtained as a graph total by defining

$$y_{ij} = \begin{cases} 2f_{ij} & \text{if } i \neq j, \\ g_{ij} + h_{ij} & \text{if } i = j. \end{cases}$$  \hspace{1cm} (65)

If the generated and the incoming flow requires sorting at each vertex, then the total sorting time can be obtained as a graph total by defining

$$y_{ij} = \begin{cases} f_{ij} & \text{if } i \neq j, \\ g_{ij} & \text{if } i = j. \end{cases}$$  \hspace{1cm} (66)

where the sorting time per flow unit is used as a time unit.

Consider an information transmission system where documents are generated at the vertices, are copied and distributed to all the reachable vertices. If $z_{ij}$ is a path indicator defined by

$$z_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or if there is a path from } i \text{ to } j, \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (67)

and if $y_{ij} = g_{ij}z_{ij}$ then the graph total becomes equal to the total number of generated and copied documents. If we introduce the reachability $z_i$, we can obtain the graph total as a vertex-value total with vertex values $g_i z_i$, for $i \in V$. Distinct information groups, i.e. subsets of vertices receiving the same documents, are represented by the connected components of the graph. If the graph is directed the groups may have overlapping information, otherwise not. The total number of distinct information groups can be obtained as a vertex value total by defining the vertex values as the inverted reachabilities $1/z_i$.

Partial observation of a flow system can be effected by taking a vertex sample $S$ and by making observations of $y_{ij}$ for $(i, j) \in S^2$ where $S^2$ can be equal to $S^2$, $S \times V$, $S \times V \cup V \times S$, etc. Another possibility is to choose a sample of flow units and observe their walks. This means that some of the path indicators $z_{ij}$ are observed. We are not, however, going to consider any kinds of walk sampling procedures in this paper.

Some important kinds of partial observation based on edge sampling can be illustrated by the following example. Assume that the graph describes some sort of incorrect or illegal economic transactions between the individuals in a population. Assume that some of the improper transactions can be discovered by a routine control of a large mass of transactions including all the improper ones. If an improper transaction from individual $i$ to individual $j$ is discovered, then a more detailed
examination of the transactions from \( i \) to \( j \) can be performed and lead to the discovering of all the improper transactions from \( i \) to \( j \), say, \( y_i = |E_{ij}| \) in number. A more thorough control procedure might examine all the outgoing transactions from \( i \) and all the incoming transactions to \( j \) and find the improper ones, i.e., \( E_i \cup E_j \). Some alternative control procedures might examine all the incoming and/or outgoing transactions at \( i \) and/or \( j \). The initially found set \( S \) of improper transactions will be the use of a combined routine and detailed control scheme give information about \( y_{ij} \) for \((i, j) \in S\) where \( S' \) can be equal to \( I(S), I_1(S) \times V \cup V \times I_2(S), [I_1(S) \cup I_2(S)] \times V \), etc.

Other kinds of graph models which are of interest in cluster analysis and multidimensional scaling are the similarity or dissimilarity structures. The vertices can be plants, bacteria, finger prints or some other kind of objects which we want to compare and separate into classes of similar objects. If object \( i \) is characterized by a data vector \((x_{i1}, \ldots, x_{in})\) for \( i = 1, \ldots, N \), then the dissimilarity between objects \( i \) and \( j \) can be defined as the squared Euclidean distance

\[
y_{ij} = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2.
\]

The corresponding graph total becomes equal to

\[
T = \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} = 2N^2 \sum_{k=1}^{n} \sigma_k^2,
\]

where

\[
\sigma_k^2 = \frac{1}{N} \sum_{i=1}^{N} (x_{ik} - \bar{x}_k)^2
\]

is the variance of the \( k \)th component of the data vectors. The graph total is an overall measure of dissimilarity. Assume that we know the data vectors of the objects in a sample \( S \). Then we can calculate the dissimilarities \( y_{ij} \) for \((i, j) \in S^2\), and consequently we can consider the problem of estimating \( T \) as a graph total estimation problem based on \( S^2 \) data.

9. Estimation of a graph total from \( S^2 \) data

Let the initial sample be a vertex sample \( S \subseteq V \) and consider the simple case of a vertex-pair sample \( S' = S^2 \) consisting of the pairs of vertices both of which belong to \( S \). Obviously \( \pi_{ij} = \pi_{ij} \) and \( \pi_{ijkl} = \pi_{ijkl} \) and by application of (57)-(60) we obtain the following theorem.

\[\textbf{Theorem 5.}\] Let \( S \) be a vertex sample with \( \pi_{ij} > 0 \) for \((i, j) \in V^2 \). Then the Horvitz-Thompson estimator of \( T \) based on \( S' = S^2 \) is given by (57) where \( \pi'_{ij} = \pi_{ij} \), and its variance is given by (58) where

\[
\gamma_{ijkl} = \frac{\pi_{ijkl} - \pi_{ij} \pi_{kl}}{\pi_{ij} \pi_{kl}}.
\]
If \( \pi_{ijkl} > 0 \) for \((i, j, k, l) \in V^4 \), then an unbiased variance estimator is given by (60) where \( \pi_{ijkl} = \pi_{ijkl} \).

It is possible to simplify the above result considerably by specifying the initial sampling design to a kind of symmetric design which has \( \pi_{ijkl} \) depending on the number but not the identities of the distinct vertices in the quadruple \((i, j, k, l)\). Simple random sampling with or without replacement and simple Bernoulli sampling are examples of such symmetric designs. Frank (1971, 1977) has given simplified versions of the variance \( \sigma^2(\hat{T}) \) and its estimator in the case of symmetric designs. In order to illustrate these results we shall now state some results pertaining to simple Bernoulli sampling.

**Corollary 6.** Let \( S \) be a vertex sample selected according to a simple Bernoulli \( \{p\} \) design with \( p = 1 - q > 0 \). Then the Horvitz–Thompson estimator of \( T \) based on \( S' = S^2 \) is given by

\[
\hat{T} = \frac{1}{p} \sum_{i \in S} y_{ii} + \frac{1}{p^2} \sum_{i \neq j} \sum_{j \in S} y_{ij},
\]

and its variance is given by

\[
\sigma^2(\hat{T}) = \frac{q}{p} \sum_{i \in V} (y_{ii} + y_{ii} - y_{ii})^2 + \frac{q^2}{2p^2} \sum_{i \neq j} \sum_{j \in S} (y_{ij} + y_{ji})^2.
\]

An unbiased estimator of the variance is given by

\[
\hat{\sigma}^2(\hat{T}) = \frac{q}{p^2} \sum_{i \in S} \left[ \frac{1}{p} \sum_{j \in S} \frac{1}{p} \sum_{j \in S} (y_{ij} + y_{ji}) \right]^2 - \frac{q^2}{2p^4} \sum_{i \neq j} \sum_{j \in S} (y_{ij} + y_{ji})^2.
\]

**Proof.** Substituting

\[
\pi'_{ij} = \pi_{ij} = \begin{cases} p & \text{if } i = j, \\ p^2 & \text{if } i \neq j \end{cases}
\]

into (57) yields \( \hat{T} \). From (59) we obtain

\[
\gamma_{ijkl} = \frac{p' - p_{r+s}}{p_{r+s}} \text{ if } (i, j, k, l) \in C_{rst}
\]

where \( C_{rst} \) is defined as the set of those \((i, j, k, l)\) which have \( r \) and \( s \) different vertices in the two pairs \((i, j)\) and \((k, l)\), and \( t \) different vertices in the quadruple \((i, j, k, l)\). The range of \((r, s, t)\) is the set

\[
J = \{(1, 1, 1), (1, 1, 2), (1, 2, 2), (1, 2, 3), (2, 2, 2), (2, 2, 3), (2, 2, 4)\}.
\]
Substituting (76) into (58) yields

$$\sigma^2(\mathcal{T}) = \sum_{i,r, s, j \in J} \frac{p'_r - p'_s}{p'_s} A_{rst}$$  \hspace{1cm} (78)$$

where

$$A_{rst} = \sum_{(i,j,k,l) \in C_{rst}} y_{ij} y_{kl}.$$ \hspace{1cm} (79)$$

The $A_{rst}$ can be given by

$$A_{111} = \sum y_{ii}, \quad A_{112} = \sum y_{ii} y_{jj}, \quad A_{122} = 2 \sum y_{ii} (y_{ij} + y_{ji}),$$

$$A_{123} = \sum y_{kk} (y_{ij} + y_{ji}), \quad A_{222} = \sum y_{ij} (y_{i} + y_{j}),$$

$$A_{223} = \sum (y_{ij} y_{kl} + 2 y_{ij} y_{kl} + y_{jk} y_{ki}), \quad A_{224} = \sum y_{ij} y_{kl},$$ \hspace{1cm} (80)$$

where all the sums are taken with different values in $V$ for all the indices. Using a number of lengthy algebraic calculations it is possible to show that (78) can be simplified to (73). Finally, substituting

$$\pi'_{ijkl} = \pi_{ijkl} = p' \quad \text{if} \quad (i, j, k, l) \in C_{rst}$$ \hspace{1cm} (81)$$

into (60) and simplifying yields (74). This completes the proof.

In particular this corollary can be applied to obtain an estimator of the edge frequency of a general graph, if the sample data consist of the subgraph generated by the initial vertex sample $S$. The following corollary gives the result for a simple undirected graph.

**Corollary 7.** Let $R$ be the total edge frequency and $Q$ the sum of squares of the local edge frequencies of the vertices in a simple undirected graph. Let $R(S)$ and $Q(S)$ be the same entities in the subgraph generated by a vertex sample $S$ selected according to a simple Bernoulli ($p$) design with $p = 1 - q > 0$. Then $R$ has an unbiased estimator

$$\hat{R} = \frac{R(S)}{p^2}$$ \hspace{1cm} (82)$$

which has the variance

$$\sigma^2(\hat{R}) = \frac{q}{p} Q + \frac{q^2}{p^2} R,$$ \hspace{1cm} (83)$$
and an unbiased variance estimator is given by

$$\delta^2(\hat{R}) = \frac{q}{p^4} Q(S) - \frac{q^2}{p^4} R(S).$$  \hfill (84)

**Proof.** If $y_{ij} = x_{ij}$ is the incidence indicator, then

$$T = 2R = \sum_{i \in V} \sum_{j \in V} y_{ij},$$  \hfill (85)

$$Q = \sum_{i \in V} \left( \sum_{j \in V} y_{ij} \right)^2,$$

and it follows from (80) that

$$A_{111} = A_{112} = A_{122} = A_{123} = 0,$$
$$A_{222} = 4R, \quad A_{223} = 4Q - 8R,$$
$$A_{224} = 4R^2 - 4Q + 4R.$$  \hfill (86)

Applying Corollary 6 to $\hat{T} = 2\hat{R}$ yields the result.

## 10. Estimation of a graph total from $S \times V \cup V \times S$ data

We shall now consider sample data $y_{ij}$ for $(i, j) \in S'$ where $S' = S \times V \cup V \times S$, i.e. the vertex-pair sample $S'$ consists of the pairs of vertices of which at least one belongs to the initial sample $S$.

**Theorem 8.** Let $S$ be a vertex sample with $\pi_{ij} < 1$ for $(i, j) \in V^2$. Then the Horvitz–Thompson estimator of $T$ based on $S' = S \times V \cup V \times S$ is given by (57) where $\pi'_{ij} = 1 - \pi_{ij}$, and its variance is given by (58) where

$$\gamma_{ijkl} = \frac{\pi_{ijkl} - \pi_{ij} \pi_{kl}}{(1 - \pi_{ij})(1 - \pi_{kl})}.$$  \hfill (87)

If $\pi_{ij} + \pi_{kl} - \pi_{ijkl} < 1$ for $(i, j, k, l) \in V^4$, then an unbiased variance estimator is given by (60) where $\pi'_{ijkl} = 1 - \pi_{ij} - \pi_{kl} + \pi_{ijkl}$.

**Proof.** If $S'_{ij}$ and $S_i$ are inclusion indicators of $S'$ and $S$, then it can easily be seen that

$$S'_{ij} = 1 - (1 - S_i)(1 - S_j).$$  \hfill (88)
and it follows that \( \pi_{ij} = \bar{\pi}_{ij} \) and \( \bar{\pi}_{ijkl} = \bar{\pi}_{ijkl} \). Consequently

\[
\pi_{ij} = 1 - \bar{\pi}_{ij} = 1 - \bar{\pi}_{ij}.
\] (89)

By applying (20) to \( S \) it follows that

\[
\pi_{ijkl} = 1 - \bar{\pi}_{ij} - \bar{\pi}_{kl} + \bar{\pi}_{ijkl} = 1 - \bar{\pi}_{ij} - \bar{\pi}_{kl} + \bar{\pi}_{ijkl}
\] (90)

and the expression for \( \gamma_{ijkl} \) then follows by substitution.

**Corollary 9.** Let \( S \) be a vertex sample selected according to a simple Bernoulli \( p \) design with \( p = 1 - q > 0 \). Then the Horvitz–Thompson estimator of \( T \) based on \( S' = S \times V \cup V \times S \) is given by

\[
\hat{T} = \frac{1}{p} \sum_{i \in S} y_i + \frac{1}{1 - q^2} \left[ \sum_{i \in S} \sum_{j \in S} (y_{ij} + y_{ji}) - \sum_{i \in S} \sum_{l \in S} y_{ij} \right],
\] (91)

and its variance is given by

\[
\sigma^2(\hat{T}) = \frac{q}{p} A_{111} + \frac{q^2}{1 - q^2} (A_{122} + A_{222}) + \frac{q^3}{p(1 + q)^2} A_{223}
\] (92)

where the \( A_{rst} \) are given by (80). An unbiased estimator of the variance is given by

\[
\hat{\sigma}^2(\hat{T}) = \frac{q}{p^2} A_{111}(S) + \frac{q^2}{p^2(1 - q^2)} \left[ A_{122}(S) + A_{222}(S) \right] + \frac{q^3}{p^4(1 + q)^2} A_{223}(S)
\] (93)

where the \( A_{rst}(S) \) are given by (80) with \( V \) replaced by \( S \).

**Proof.** Substituting

\[
\pi'_{ij} = 1 - \bar{\pi}_{ij} = \begin{cases} p & \text{if } i = j, \\ 1 - q^2 & \text{if } i \neq j \end{cases}
\] (94)

in (57) yields (91), and substituting

\[
\gamma_{ijkl} = \frac{q' - q'^{++}}{(1 - q')(1 - q^2)} \text{ if } (i, j, k, l) \in C_{rst}
\] (95)

in (58) yields (92). Here \( C_{rst} \) is defined as in (76). The Horvitz–Thompson variance estimator could be obtained by substituting

\[
\pi^*_{ijkl} = 1 - q^* - q^* + q^* \text{ if } (i, j, k, l) \in C_{rst}
\] (96)
in (60), but the simpler variance estimator given by (93) is obtained from (92) by using
that $A_{rs}^*$ has the unbiased estimator $A_{rs}^*(S)/p'$. 

In particular, Corollary 9 can be applied to obtain an estimator of the edge frequency
of a general graph, if the sample data consist of the observations of all the edges which
are incident with any of the vertices in an initial vertex sample S. The following
corollary gives the result for a simple undirected graph. The proof is straightforward
and is omitted.

**Corollary 10.** Let $R$ be the total edge frequency and $Q$ the sum of squares of the local
edge frequencies of the vertices in a simple undirected graph. Let $R'(S)$ and $R(S)$ be the
numbers of edges which are incident with only one vertex in $S$ and with two vertices in $S$,
respectively, where $S$ is a vertex sample selected according to a simple Bernoulli ($p$)
design with $p = 1 - q > 0$. Then $R$ has an unbiased estimator

$$
\hat{R} = \frac{R(S) + R'(S)}{1 - q^2}
$$

which has the variance

$$
\sigma^2(\hat{R}) = \frac{q^2}{(1 + q)^2} R + \frac{q^3}{p(1 + q)^2} Q
$$

An unbiased variance estimator is given by

$$
\hat{\sigma}^2(\hat{R}) = \frac{(p^2 - 2q^2)q^2}{p^2(1 - q^2)^2} R(S) + \frac{q^3}{p^2(1 - q^2)^2} Q(S).
$$

where $Q(S)$ is the sum of squares of the local edge frequencies of the vertices in the
subgraph generated by $S$.

A comparison between Corollaries 7 and 10 shows that the variance reduction is

$$
\frac{q(1 + 2q)}{p(1 + q)^2} R + \frac{4q^3}{p^2(1 + q)^2} Q
$$

due to the inclusion of the sample information about the edges which are incident with
only one of the sample vertices.

If the vertex pair values $y_{ij}$ are available for $S = S \times V \cup V \times S$ it is possible to observe
the vertex value

$$
y_i = \sum_{j \in V} \frac{y_{ij} + y_{ji}}{2}
$$
for each sample vertex \( i \in S \), and the graph total \( T \) can be given as a vertex-value total according to \( T = \sum y_i \). Consequently the Horvitz–Thompson estimator of this vertex-value total provides an alternative to the estimator in Theorem 8. The following theorem shows that generally this alternative neither dominates nor is dominated by the estimator in Theorem 8.

**Theorem 11.** Let \( S \) be a vertex sample having a simple Bernoulli \((p)\) design with \( p = 1 - q > 0 \). Then an unbiased estimator of \( T \) based on \( S' = S \times V \cup V \times S \) is given by

\[
T^* = \frac{1}{2p} \sum_{i \in S} \sum_{j \in V} (y_{ij} + y_{ji}).
\]

If \( y \) is a non-negative or non-positive variable, then \( T^* \) is dominated by the estimator \( \hat{T} \) in Corollary 9; otherwise the variance of \( T^* \) can be smaller than the variance of \( \hat{T} \).

**Proof.** The variance of \( T^* \) is equal to

\[
\sigma^2(T^*) = \frac{q}{4p} \sum_{i \in V} \left[ \sum_{j \in V} (y_{ij} + y_{ji}) \right]^2
\]

and using (80) and Corollary 9 it follows that

\[
\sigma^2(T^*) - \sigma^2(\hat{T}) = \frac{q}{2(1+q)} (A_{122} + A_{222}) + \frac{q(1+3q)}{4(1+q)^2} A_{233}.
\]

According to (80) this difference is non-negative if \( y_{ij} \geq 0 \) (or \( y_{ij} \leq 0 \)) for all \((i, j) \in V^2\). The difference becomes negative if there is a pair \((i, j)\) such that \( y_{ij} + y_{ji} > 0 \) and \( y_{ji} \) is negative and of sufficiently large absolute value.

**Corollary 12.** With the assumptions and notations in Corollary 10 \( R \) has an unbiased estimator

\[
R^* = \frac{2R(S) + R'(S)}{2p}
\]

which has a variance satisfying

\[
\sigma^2(R^*) = \frac{q}{4p} Q \geq \sigma^2(\hat{R}) + \frac{q}{2(1+q)} \frac{R}{R'}
\]

with equality if and only if the graph has at most one incident edge at each vertex.
Proof. The result follows from (85), (86), (103) and (104) by noticing that

$$Q - 2R = \sum_{i \in V} \left[ \left( \sum_{j \in V} y_{ij} \right)^2 - \sum_{j \in V} y_{ij} \right] \geq 0$$  \hspace{1cm} (107)

with equality if and only if $\sum j y_{ij}$ is equal to 0 or 1 for each $i \in V$.

11. Estimation of a graph total from some other kinds of vertex-sample generated data

The general formulae in Section 7 can be applied to many other kinds of data than the two kinds considered in Sections 9 and 10. We shall illustrate this by giving some further examples of vertex-sample generated data.

We shall first consider sample data $y_{ij}$ for $(i, j) \in S'$ where $S'$ is equal to $S \times V$ or $S \times A(S)$. The inclusion indicators of $S'$ become

$$S'_{ij} = S_i,$$  \hspace{1cm} (108)

$$S'_{ij} = S_i \left[ 1 - \prod_{k \in V} (1 - S_{k,x_{ij}}) \right]$$  \hspace{1cm} (109)

in the two cases, respectively. In the first case the Horvitz–Thompson estimator of the graph total $T$ becomes

$$T = \sum_{i \in S} \sum_{j \in V} \frac{y_{ij}}{\pi_i}$$  \hspace{1cm} (110)

and the formulae (57)-(60) can be directly applied. In the other case it follows after a certain amount of algebra that

$$\pi'_{ij} = \pi_i + \pi(B_j) - \pi(i \cup B_j),$$  \hspace{1cm} (111)

$$\pi'_{ijk} = \pi_{ik} + \pi(i \cup B_j) + \pi(k \cup B_j) + \pi(B_j \cup B_i) - \pi(i \cup B_j \cup B_i) - \pi(i \cup k \cup B_j) - \pi(k \cup B_i) - \pi(i \cup B_i) - \pi(k \cup B_j)$$  \hspace{1cm} (112)

and

$$\hat{T} = \sum_{i \in S} \sum_{j \in A(S)} \frac{y_{ij}}{\pi_i - \pi(B_j) + \pi(i \cup B_j)},$$  \hspace{1cm} (113)

and (57)-(60) can be applied. In particular if $x_{ij}=0$ implies that $y_{ij}=0$, then the two estimators (110) and (113) become identical.

Consider now the case of sample data from all the pairs in a one-stage snowball
The application of the variance formula (58) is somewhat simplified if \( x_{ij} = 0 \) implies \( \gamma_{ij} = 0 \). We shall now consider such a case. If we define

\[
v_{ij} = \begin{cases} x_{ij} x_{ji} & \text{for } i \neq j, \\ 0 & \text{otherwise}, \end{cases}
\]

then the total \( T \) becomes equal to the total number of mutual edges in the graph. The following theorem gives the estimator of the total mutual edge frequency pertaining to a simple Bernoulli (\( p \)) design and a graph with exactly one loop and one other outedge at each vertex.

**Theorem 13.** Let \( S \) be a vertex sample having a simple Bernoulli (\( p \)) design with \( p = 1 - q > 0 \). Assume that \( x_{ii} = 1 \) and \( x_{i} = 2 \) for \( i \in V \). Then an unbiased estimator of the total number of mutual edges.

\[
T = \sum_{i \neq j} x_{ij} x_{ji},
\]

based on \( S' = A(S)^2 \) is given by

\[
T = \sum_{(i,j) \in S} \frac{x_{ij} x_{ji}}{1 - q^{x_{i} - 1} - q^{x_{j} - 1} + q^{x_{i} + x_{j} - 2}},
\]

and its variance is given by

\[
\sigma^2(T) = \frac{1 - q^4}{q^4} \left[ \sum_{i \neq j} \sum_{(i,j) \in S} \frac{x_{ij} x_{ji} q^{x_{i} + x_{j}}}{1 - q^{x_{i} - 1} - q^{x_{j} - 1} + q^{x_{i} + x_{j} - 2}} \right]^2.
\]
Note. It should be noticed that the estimator requires the observability of the total number of inedges at every vertex in the snowball sample $A(S)$. If $M_{rs}$ and $M_{rs}(S')$ denote the numbers of mutual edges between the unordered vertex pairs in $V^2$ and $S'$ which have $r$ and $s$ inedges, then (118) and (119) can be given by

$$
\hat{T} = \sum_{r \leq s} \sum_{1 - q^r - q^s + q^{r+s-2}} \frac{M_{rs}(S')}{1 - q^r - q^s + q^{r+s-2}}, \quad (120)
$$

$$
\sigma^2(\hat{T}) = \frac{1}{q^4} \left[ \sum_{r \leq s} \sum_{1 - q^r - q^s + q^{r+s-2}} \frac{M_{rs}q^{r+s}}{1 - q^r - q^s + q^{r+s-2}} \right]^2. \quad (121)
$$

Proof. According to (115) a simple Bernoulli ($p$) design implies that

$$
\pi'_{ij} = 1 - q^{x_i} - q^{x_j} + q^{x_i + x_j} - x_{ij} - x_{ji} \quad (122)
$$

and (118) follows. Moreover, assuming

$$
x_{ij} x_{ji} = x_{kl} x_{lk} = 1, \quad (123)
$$

it follows from (115) and (116) that

$$
\pi'_{ijk} - \pi'_{ij} \pi'_{kl} = (1 - q^4)q^{x_i + x_j + x_k + x_l} - 4. \quad (124)
$$

According to (20), (58), (59) and (124) we obtain (119) by substitution and simplification.

12. Estimation of a graph total from some different kinds of edge-sample generated data

We shall now turn to vertex-pair samples $S \subseteq V^2$ generated by initial edge samples $S \subseteq E$.

Consider first a directed graph and define $S'$ as the set of the incident vertex-pairs of the edges in $S$, i.e. $S' = I(S)$. Then the sample inclusion indicators satisfy

$$
S'_{ij} = \begin{cases} 
1 - \prod_{\omega} (1 - S_\omega) & \text{if } E_{ij} \neq \emptyset, \\
0 & \text{otherwise},
\end{cases} \quad (125)
$$

and we obtain

$$
\hat{\pi}'(M) = \hat{\pi}' \left( \bigcup_{(i,j) \in M} E_{ij} \right) \text{ for } M \subseteq V^2. \quad (126)
$$

Hence we have the following theorem.
Theorem 14. Let $S$ be an edge sample and $S' = I(S)$. The Horvitz-Thompson estimator of $T$ based on $S'$ is then given by (57) if $\pi_{ij} = 1 - \tilde{\pi}(E_{ij}) > 0$ unless $y_{ij} = 0$. Its variance is given by (58) and (72) where

$$\pi_{ijk} = 1 - \pi(E_{ij}) - \pi(E_{ki}) + \pi(E_{ij} \cup E_{ki}).$$

An unbiased variance estimator is given by (60) if $\pi_{ijk} > 0$ unless $y_{ij}y_{ki} = 0$.

If $S$ has a Bernoulli design it follows from (125) that $S'$ has a Bernoulli design too, and we obtain the simple formulae

$$\hat{T} = \sum_{i,j \in S} \frac{y_{ij}}{1 - \pi(E_{ij})},$$

$$\sigma^2(\hat{T}) = \sum_{i,j \in S} \frac{y_{ij}^2 \tilde{\pi}(E_{ij})}{1 - \pi(E_{ij})},$$

$$\hat{\sigma}^2(\hat{T}) = \sum_{i,j \in S} \frac{y_{ij}^2 \tilde{\pi}(E_{ij})}{1 - \pi(E_{ij})}.\quad (130)$$

In order to illustrate these formulae we consider a simple particular case. Assume that one observes the subgraph generated by $I_1(S) \cup I_2(S)$ in a general graph, i.e. the subgraph generated by $S \subseteq E$ in a simple graph. Put $y_{ij} = |E_{ij}|$ which implies that $T$ becomes equal to the edge frequency in a directed graph. If $S$ has a simple Bernoulli ($p$) design with $p = 1 - q > 0$ the estimator becomes equal to

$$\hat{T} = \sum_{r \geq 1} \frac{m_r(S')}{1 - q^r},\quad (131)$$

where $m_r(S')$ is the number of ordered vertex pairs in $S'$ which are incident with exactly $r$ parallel edges. If $m_r$ is the total number of ordered vertex pairs which are incident with exactly $r$ parallel edges then the variance of $\hat{T}$ is given by

$$\sigma^2(\hat{T}) = \sum_{r \geq 1} \frac{r^2 q^r m_r}{1 - q^r},\quad (132)$$

and an unbiased variance estimator is given by

$$\hat{\sigma}^2(\hat{T}) = \sum_{r \geq 1} \frac{r^2 q^r m_r(S')}{(1 - q^r)^2}.$$  

A simple direct proof is easily obtained by using the fact that $m_r(S')$ is binomially distributed with the parameters $m_r$ and $1 - q^r$. 

Survey sampling in graphs 261
Theorem 15. Let $S$ be an edge sample and $S' = I_1(S) \times V \cup V \times I_2(S)$. The Horvitz-Thompson estimator of $T$ based on $S'$ is then given by (57) if $\pi_{ij}' = 1 - \pi(E_i \cup E_j) > 0$ unless $y_{ij} = 0$. Its variance is given by (58) and (59) where

$$\pi_{ijkl}' = 1 - \pi(E_i \cup E_j) - \pi(E_k \cup E_1) + \pi(E_i \cup E_j \cup E_k \cup E_1).$$

(134)

An unbiased variance estimator is given by (60) if $\pi_{ijkl}' > 0$ unless $y_{ij}y_{kl} = 0$.

Proof. The sample inclusion indicators satisfy

$$S_{ij}' = \begin{cases} 1 - \prod_{e \in E_i \cup E_j} (1 - Se) & \text{if } E_i \cup E_j \neq \emptyset, \\ 0 & \text{otherwise}, \end{cases}$$

(135)

and it follows that

$$\tilde{\pi}'(M) = \tilde{\pi}\left( \bigcup_{(i,j) \in M} (E_i \cup E_j) \right) \text{ for } M \subseteq V^2.$$

(136)

An application of (20) will complete the proof.

In particular we obtain for a Bernoulli design that $\gamma_{ijkl}$ in (59) can be given by

$$\gamma_{ijkl} = \frac{\phi(E_{il} \cup E_{kj})}{\phi(E_i \cup E_j) \phi(E_k \cup E_1)}$$

(137)

where $\phi$ is defined by

$$\phi(M) = \frac{1 - \tilde{\pi}(M)}{\tilde{\pi}(M)}.$$

(138)

For a simple Bernoulli ($p$) design with $0 < p < 1$ and $q = 1 - p$ we obtain

$$\pi_{ij}' = 1 - q^r \text{ for } (i, j) \in C_r,$$

(139)

$$\gamma_{ijkl} = \frac{q^{s+t}(1 - q^s)}{(1 - q^r)(1 - q^s)} \text{ for } (i, j, k, l) \in C_{rst},$$

(140)

$$\pi_{ijkl}' = 1 - q^r - q^s + q^{s+t} \text{ for } (i, j, k, l) \in C_{rst},$$

(141)

where

$$C_r = \{(i, j) \in V^2 : |E_i \cup E_j| = r\},$$

(142)

$$C_{rst} = \{(i, j, k, l) \in V^4 : |E_i \cup E_j| = r, |E_k \cup E_1| = s\}.$$  

(143)
Survey sampling in graphs

If we now introduce the parameters

\[ Y_{rst} = \sum_{(i,j,k,l) \in C_{rs}} y_{ij} y_{kl} \]  

(144)

and the statistics

\[ Y_r(S') = \sum_{(i,j) \in S' \cap C_s} y_{ij} \]  

(145)

\[ Y_{rst}(S') = \sum_{(i,j,k,l) \in (S')^2 \cap C_{rs}} y_{ij} y_{kl} \]  

(146)

we obtain

\[ \hat{r} = \frac{\sum_{r \geq 1} Y_r(S')}{1 - q^r}, \]  

(147)

\[ \sigma^2(\hat{r}) = \sum_{r,s,t} \frac{q^{r+s-t}(1-q^r)}{(1-q^r)(1-q^s)} Y_{rst}, \]  

(148)

\[ \delta^2(\hat{r}) = \sum_{r,s,t} \frac{q^{r+s-t}(1-q^r)}{(1-q^r)(1-q^s)(1-q^r-q^s+q^{r+s-t})} Y_{rst}(S'). \]  

(149)

The results above can be modified and extended in various ways to cope with some related sampling procedures, for instance \( S' = I_1(S) \times I_2(S) \) and \( S' = [I_1(S) \cup I_2(S)]^2 \). The application of our tools to these cases does not present any new problems in principle, but may provide interesting results if different sampling and observation procedures are compared for various kinds of graphs.

References


