Bipartite Incidence Graph Sampling

Li-Chun Zhang* Melike Oğuz-Alper †

Abstract

Graph sampling provides a statistical approach to study real graphs, which can be of interest in numerous investigations. There have been significant contributions to the existing graph sampling theory. However, a general approach to graph sampling which also unifies the existing unconventional sampling methods, which may be envisaged as graph sampling problems, including indirect, network and adaptive cluster sampling, as well as arbitrary T-stage snowball sampling, is non-existent in the literature. We propose a bipartite incident graph sampling (BIGS) as a feasible representation of graph sampling from arbitrary finite graphs and a unified approach to a large number of graph sampling situations. We establish the sufficient and necessary conditions under which the BIGS is feasible for various graph sampling methods. Under a feasible BIGS, two types of design-unbiased estimators, the Horvitz-Thompson estimator and the Hansen-Hurwitz type of estimators, can be applied. A general result on the relative efficiency of the two types of estimators is obtained. Some numerical results based on a limited simulation study illustrating the feasibility of the proposed approach are presented.

Key Words: Graph sampling, adaptive cluster sampling, indirect sampling, network sampling, T-stage snowball sampling, ancestral observation procedure

1. Introduction

Graph sampling provides a statistical approach to study real graphs, which represent the structure of many technological, social or biological phenomena of interest. It is based on exploring the variation over all possible subsets of nodes and edges, i.e. *sample graphs*, which can be taken from the given *population graph*, according to a specified method of sampling. Zhang and Patone (2017) synthesise the existing graph sampling theory, extending the previous works on this topic by Frank (1971, 1980*a*,*b*, 2011). A general definition is given for probability sample graphs, and the unbiased Horvitz and Thompson (HT) (1952) estimator is developed for arbitrary T-stage snowball sampling (T-SBS) from finite graphs.

The unconventional sampling methods, such as, multiplicity sampling (Birnbaum and Sirken, 1965), adaptive cluster sampling (ACS) (Thompson, 1990) and indirect sampling (Lavallée, 2007) are envisaged, for the first time, as graph sampling problems by Zhang and Patone (2017). Various graph representations of these sampling methods are provided. However, a general approach to graph sampling unifying the existing unconventional sampling methods, as well as arbitrary T-SBS, is non-existent in the literature.

We propose a bipartite incident graph sampling (BIGS) representation of sampling from finite graphs as a flexible and unified approach to a large number of graph sampling situations. The main contribution of this paper is to establish the sufficient and necessary conditions under which the BIGS is *feasible* for various graph sampling methods. We derive the number of stages required for a feasible BIGS representation of an arbitrary T-SBS from a finite population graph. BIGS representation of T-SBS enables simple computation of the inclusion probabilities of sample motifs, without requiring recursive computations.

*University of Southampton, SO17 1BJ, Southampton, UK; Statistics Norway, Otervegen 23 2211, Kongsvinger, Norway; University of Oslo, Niels Henrik Abels hus, Moltke Moes vei 35, 0851, Oslo, Norway

†Statistics Norway, Akersveien 26, 0177, Oslo, Norway

The Hansen and Hurwitz (HH) (1943) type estimator has been used in many works on network sampling, as summarised by Sirken (2005). The HT estimator is presented as a method of estimation in graph sampling theory (e.g. (Frank, 1971; Zhang and Patone, 2017)). Under feasible BIGS, there are several choices of unbiased estimators applicable that allows one to explore the potentials of efficiency gains in graph sampling. We provide a general result on the relative efficiency of the HT and the HH type of estimators, which is unknown in the literature discussing both estimators (e.g. Thompson, 1990, 1991, 2012).

The rest of the paper is organised as follows. Graph sampling is described in Section 2. BIG sampling is presented in Section 3, followed by Section 3.1 in which the sufficient and necessary conditions for a feasible BIGS representation of graph sampling are established. In Section 4, formal BIGS representations are described for the aforementioned unconventional sampling methods. In Section 5, we develop the BIGS representation of general T-stage snowball sampling, including the relevant results for identifying the sample motifs eligible for estimation. In Section 6, the general condition governing the relative efficiency of the HT and HH-type estimators under BIG sampling is presented. In Section 7, some numerical results are provided for an example of T-stage snowball sampling from an arbitrary population graph. Finally, some concluding remarks are given in Section 8.

2. Graph sampling

Let $G = (U, A)$ be the population graph, with node set U and edge set A. We focus on simple graphs in this paper for simplicity, such that there can be at most one edge between a pair of nodes (i, j), where $i, j \in U$. Let $a_{ij} = 1$ if edge $(ij) \in A$ and 0 otherwise. By definition $a_{ji} \neq a_{ij}$ if the graph is *directed*, but $a_{ij} \equiv a_{ij}$ if the graph is *undirected*. The theory developed below can be easily adapted to multigraphs, where there can be more than one edge between any pair of nodes.

The measurement units of interest are called the *motifs* in G. Denote by $\Omega = \Omega(G)$ the set of all motifs in G. For any $k \in \Omega$, let M_k be the nodes involved in the motif k, of *order* $|M_k|$. The motif of these M_k nodes is denoted by $[M_k]$, such that for any $k \neq l \in \Omega$, we have $[M_k] = [M_l]$, but $M_k \neq M_l$, nor is it necessary that $|M_k| = |M_l|$.

Zhang and Patone (2017) give the following general definition of sample graphs from G. Let s_0 be an initial sample of nodes taken from the sampling frame F, where $s_0 \subset F \subseteq U$, according to the sampling distribution $p(s_0)$, and $\sum_{s_0} p(s_0) = 1$ and $\pi_i = \Pr(i \in s_0) > 0$ for any $i \in F$. Given s_0, graph sampling proceeds according to a specified *observation procedure (OP)*, for edges that are *incident* to the nodes in s_0. The observed edges, denoted by A_s for $A_s \subseteq A$, are specified using a *reference set* s_{ref}, where $s_{ref} \subseteq U \times U$, such that any existing edge (ij) in A is observed if $(ij) \in s_{ref}$. That is, s_{ref} specifies the parts of the adjacency matrix that are observed under the given OP. Denote by $\text{Inc}(a_{ij}) = \{i, j\}$ the nodes that are incident to the edge (ij). Let $\text{Inc}(A_s) = \cup_{a_{ij} \in A_s} \text{Inc}(a_{ij})$ be the set of nodes incident to the edges A_s. The *sample graph* is given by

$$G_s = (U_s, A_s) \quad \text{and} \quad U_s = s_0 \cup \text{Inc}(A_s) \, .$$

The motifs that are observed in the sample graph G_s can now be given as follows: $\forall k \in \Omega$, we have $k \in \Omega_s = \Omega(G_s)$, iff $M_k \times M_k \subseteq s_{ref}$. In particular, notice that $M_k \subseteq A_s$ does not imply $k \in \Omega_s$ in general, but $k \in \Omega_s$ must imply $M_k \in U_s$.

3. BIG sampling

Graph sampling can be given a BIGS representation, provided the following. Let

$$\mathcal{B} = (F \cup \Omega; H)$$

be the BIG associated with the population graph G and the motif set $\Omega = \Omega(G)$, where the edges H exist only between F and Ω but not between any $i, j \in F$ or $k, l \in \Omega$, and an edge exists from any $i \in F$ to $k \in \Omega$ iff $k \in \Omega_s$ whenever $i \in s_0$, so that graph sampling from G can be represented as sampling from $\mathcal{B}$ by *incident* OP (Zhang and Patone, 2017) given s_0. This clarifies the term *bipartite incidence graph*.

For any graph sampling from given G, let $\delta_{i,k} = 1$ for any $i \in F$ and $k \in \Omega$, iff $k \in \Omega_s$ whenever $i \in s_0$, or $\Pr(k \in \Omega_s | i \in s_0) = 1$, according to the graph sampling design, which consists of $p(s_0)$ and the OP given s_0. For any $i \in F$, let

$$\alpha_i = \{k : k \in \Omega, \delta_{i,k} = 1\} ,$$

which contains all the *successors* of i in $\mathcal{B}$; for any $k \in \Omega$, let

$$\beta_k = \{i : i \in F, \delta_{i,k} = 1\} ,$$

which contains all the predecessors of k in $\mathcal{B}$. In other words, $(ik) \in H$ or $h_{ik} = 1$ in $\mathcal{B}$, iff $\delta_{i,k} = 1$ for $i \in F$ and $k \in \Omega$. The *sample BIG* is given by

$$\mathcal{B}_s = (s_0, \Omega_s; H_s) \quad \text{and} \quad \Omega_s = \alpha(s_0) = \cup_{i \in s_0} \alpha_i \quad \text{and} \quad H_s = H \cap (s \times \Omega_s) . \qquad (1)$$

3.1 The sufficient and necessary conditions

The feasibility of a BIGS representation of graph sampling from G can be determined based on the sufficient and necessary conditions given in Theorem 1

Theorem 1. *Graph sampling from $G = (U, A)$ with associated motifs Ω of interest, based on $p(s_0)$ and the given OP, can be represented by ancestral BIG sampling from $\mathcal{B}$, iff*

(i) $\forall k \in \Omega$ and $i \in F$, $\delta_{i,k} = 1$ or 0 in G can be determined given $i \in s_0$ alone;

(ii) $\forall k \in \Omega$, we have $\beta_k \neq \emptyset$ in $\mathcal{B}$, or equivalently $\underset{i \in F}{\cup} \alpha_i = \Omega$ in $\mathcal{B}$;

(iii) graph sampling OP in G ensures the observation of $\beta\big(\alpha(s_0)\big) \setminus s_0$ in $\mathcal{B}$.

Given (i), we can define the edge set H of $\mathcal{B} = (F, \Omega; H)$. Given (ii), BIG sampling covers all the motifs in Ω, since $\Pr(k \in \Omega_s)$ is then positive for any $k \in \Omega$. Given (iii), it is possible to calculate the inclusion probability of $k \in \Omega_s$, based on $p(s_0)$ for $s_0 \subset F$. Thus, conditions (i) - (iii) are sufficient. They are also necessary, because removing any of them would render the BIGS representation infeasible.

4. BIGS representation for unconventional sampling methods

Below we describe formally BIGS representation as a unified approach to indirect sampling, network sampling and ACS.

4.1 Indirect sampling

Generally for indirect sampling, let F be the sampling frame, and Ω the set of measurement units of interest, which are accessible via the sampling units in F. For instance, F can be the hospitals and Ω the patients treated by the hospitals in F (e.g. Birnbaum and Sirken, 1965). Analogically, F can be all the parents and Ω the children to the people in F (e.g. Lavallée, 2007). For any $i \in F$ and $k \in \Omega$, we have $(ik) \in H$ or $h_{ik} = 1$ iff k can be reached given $i \in s_0$, denoted by $\delta_{i,k} = 1$. This completes the definition of population graph $\mathcal{B} = (F, \Omega; H)$. The knowledge of multiplicity that is collected under indirect sampling

ensures then ancestral BIG sampling from $s_0 \subset F$, where the sample BIG is given by (1), with the associated out-of-sample *ancestors* $\beta\big(\alpha(s_0)\big) \setminus s_0$ in $\mathcal{B}$.

The probability of inclusion in Ω_s can be derived from the initial sampling distribution $p(s_0)$, for $s_0 \subset F$. The (first-order) inclusion probability of $k \in \Omega_s$ is given by

$$\pi_{(k)} = 1 - \bar{\pi}_{\beta_k} = 1 - \Pr\big(\cap_{i \in \beta_k} i \notin s_0 \big) \,, \tag{2}$$

where $\bar{\pi}_{\beta_k}$ is the exclusion probability of β_k in s_0, i.e. the probability that none of the ancestors of k in $\mathcal{B}$ is included in the initial sample s_0. Notice that the knowledge of the out-of-sample ancestors $\beta_k \setminus s_0$ is required to compute $\bar{\pi}_{\beta_k}$. Similarly, the second-order inclusion probabilities of $k \neq l \in \Omega_s$ is given by

$$\pi_{(kl)} = 1 - \big(\bar{\pi}_{\beta_k} + \bar{\pi}_{\beta_l} - \bar{\pi}_{\beta_k \cup \beta_l} \big) \,. \tag{3}$$

4.2 Network sampling

Sampling of siblings via an initial sample of households provides an example of network sampling (Sirken, 2005). Since the siblings may belong to different households, some of which are outside of the initial sample, the *network* relationship among the siblings is needed. Network sampling as such can be viewed as a form of indirect sampling, since the sampling unit (household) is not the unit of measurement (siblings), and the latter cannot be sampled directly. Notice that the term *network* has a specific meaning here, unlike when network refers to a whole *valued graph* (Frank, 1980*a*,*b*), e.g. an electricity network, where the nodes and edges have associated values that are of interest.

Let F denote the sampling frame, which is the list of households from which the initial sample s_0 can be selected. Provided the OP under network sampling is *exhaustive*, in the sense that all the siblings are observed, if at least one of them belongs to a household in s_0, one can treat each network of siblings as a motif of interest, such that Ω consists of all the networks of siblings. For any $i \in F$ and $k \in \Omega$, let $(ik) \in H$ iff at least one of the siblings in M_k belongs to household i. This yields the population graph $\mathcal{B} = (F, \Omega; H)$. Network sampling with observation of multiplicity is then equivalent to ancestral BIG sampling in $\mathcal{B}$, where $\Omega_s = \alpha(s_0)$, with the associated out-of-sample *ancestors* $\beta\big(\alpha(s_0)\big) \setminus s_0$, such that the inclusion probabilities of the motifs can be calculated by (2) and (3).

4.3 Adaptive cluster sampling (ACS)

As a standard example of ACS (Thompson, 1990), let F consist of a set of spatial grids over a given area. Let y_i be the amount of a species, which can be found in the i-th grid. Given $i \in s_0$, one would survey all its neighbour grids (in four directions) if y_i exceeds a threshold value but not otherwise. The OP is repeated for all the neighbour grids, which may or may not generate further grids to be surveyed. The process is terminated, when the last observed grids are all below the threshold. The interest is to estimate the total amount of species (or mean per grid) over the given area.

One can consider each cluster of contiguous grids, where the associated y_i's all exceed the threshold value, as a network. Let a grid with y_i below the threshold value form a singleton network consisting only of itself. The OP is network exhaustive, since all the grids in a network are observed if at least one of them is selected in s_0. A singleton network is an *edge* grid, if it is contiguous to a non-singleton network. Observing a non-singleton network will lead one to observe all its edge grids, but not the other way around.

The ACS can be represented as BIG sampling from $\mathcal{B}$, where the grids are both the sampling units of F and the motifs of Ω. Let $h_{ik} = 1$ if k is observed under ACS whenever $i \in s_0$, for $i \in F$ and $k \in \Omega$. However, the OP of ACS is not ancestral when an edge grid is

selected in s_0, but none of the grids in its non-singleton neighbour network (NNN) is in s_0. In this case, one would not observe its ancestors, i.e. its NNN, in this $\mathcal{B}$ and its inclusion probability cannot be calculated based on the observed sample.

Thompson (1990) proposes the idea of *eligibility* of edge grids for estimation. An edge grid is *eligible* only if it is selected in s_0 directly, in the case of which its inclusion probability is known, but not when it is observed via its NNN. The corresponding estimator is called the modified HT estimator. Under a feasible BIGS representation, one can set $h_{ik} = 0$ in *restricted* $\mathcal{B}$ denoted by $\mathcal{B}^*$, where grid i belongs to the NNN of the edge grid k, such that k is eligible for estimation only when it is selected in s_0 directly. The unmodified HT estimator can be used under this BIGS representation. The same estimates are obtained under either of these strategies. However, the Rao-Blackwellised version of the unmodified HT estimator is unchanged, unlike that of the modified HT estimator, which differs generally from the corresponding original estimator.

5. T-stage snowball sampling (T-SBS)

Goodman (1961) considers *snowball sampling (SBS)* on a special directed graph, where each node has one and only one out-edge. Frank (1977) and Frank and Snijders (1994) consider one-stage SBS from arbitrary population graphs. Zhang and Patone (2017) derive the HT-estimator for general T-stage snowball sampling (T-SBS). Additional stages of sampling are generally needed in order to identify the ancestors of all the motifs observed under T-SBS though.

Let $G = (U, A)$ be an undirected simple graph. Let s_0 be the initial sample of *seeds* taken from $F = U$, according to $p(s_0)$, where $s_0 \subset U$. Let $s_1 = s_0 \cup \alpha(s_0)$ be the sample of nodes observed after the first stage given an OP, where $s_1 \setminus s_0$ is the first-wave snowball sample, which are the *seeds* for the second stage snowball sample, and so on. Denote by s_T the observed sample of nodes after T stages, by which time s_T may have only covered a part of a network.

For any population graph G, a motif in $\Omega(G)$ may be unobserved under T-SBS, even though it is observable under SBS with an infinite number of stages. Moreover, not all the observed motifs after T stages are eligible for estimation, and additional stages of sampling may be required in order to observe all the ancestors that could have led to an observed motif by T-SBS. However, more motifs of interest may be observed during the additional sampling, which again may or may not be eligible for estimation.

Let ν_{ij} be the length of the *geodesic* from i to j in G, which is the shortest path from i to j in G, for any $k \in \Omega$ and $i \neq j \in M_k$. Since the shortest path from i to j varies with the OP, let us assume incident reciprocal observation for simplicity. For any $k \in \Omega$ and $i \in M_k$, let $d_{i,k}$ be the SBS *observation distance* from i to k, which is the minimum number of stages required to observe $k \in \Omega_s$ under SBS from G, when starting from i. Lemma 1 defines the observation distance $d_{i,k}$.

Lemma 1. *$\forall k \in \Omega$ and $i \in M_k$, if the nodes M_k are connected in G, then*

$$d_{i,k} = \begin{cases} \max_{j \in M_k} \nu_{ij} & \textit{if } |\arg\max_{j \in M_k} \nu_{ij}| = 1 \\ 1 + \max_{j \in M_k} \nu_{ij} & \textit{otherwise} \end{cases},$$

or if there exists a single node other than i which is unconnected to i in G, then

$$d_{i,k} = 1 + \max_{j \in M_{k;i}} \nu_{ij}$$

where $M_{k;i}$ consists of the nodes in M_k that are connect to i in G.

Corollary 1. *If there exists $k \in \Omega$, where there are at least two nodes, $i \neq j \in M_k$, such that $d_{i,k} = d_{j,k} = \infty$ in G, then BIGS representation of T-SBS from G is infeasible.*

5.1 BIGS representation using all the motifs observed under T-SBS

The geodesic-distance matrix based on the sample graph G_s is generally not the same as that of the population graph G. Additional sampling in G is then necessary, in order to identify the ancestors of any observed motif in Ω_s, as specified below.

Lemma 2. *For any $k \in \Omega_s$, if $|M_k| > 1$ then one needs at most $T-1$ stages of additional SBS from M_k to observe all the ancestors of sample motif k under T-SBS from G, if $|M_k| = 1$ then one needs at most T stages of additional SBS from M_k.*

Suppose T-SBS from G is a probability sampling design for $\Omega(G)$ that is of interest. For BIGS representation of T-SBS from G, let $F = U$ and $\Omega = \Omega(G)$. By Theorem 1, one needs to set $h_{ik} = 1$ for any i that is the ancestor of motif k under T-SBS from G. One can set $h_{ik} = 1$ in the sample graph $\mathcal{B}_s$ directly, provided $k \in \Omega_s$ can be observed in G_s starting from $i \in s_0$. Moreover, having identified all the ancestors of each observed motif $k \in \Omega_s$ by additional sampling, as guaranteed under Lemma 2, one can set $h_{ik} = 1$ for all the out-of-s_0 ancestors of k under T-SBS from G. In this way, ancestral observation is achieved for all the motifs in Ω_s, such that they all can be used for estimation.

5.2 BIGS representation for eligible motifs under T-SBS

Here, we present strategies of BIGS representation that are feasible based on the eligible motifs observed under T-SBS, without additional sampling for ineligible motifs. Let $\mathcal{B}$ be the population BIG representing T-SBS from G, where all the observed motifs can be used for estimation. For each $k \in \Omega$ with ancestors β_k in $\mathcal{B}$, let β_k^* be a non-empty subset of β_k, where $\emptyset \neq \beta_k^* \subseteq \beta_k$. Consider BIG sampling with *restricted ancestors* from $\mathcal{B}^* = (F, \Omega; H^*)$, where H^* contains only the edges from β_k^* to k, for each $k \in \Omega$. Since β_k^* is non-empty for every $k \in \Omega$, conditions (i) and (ii) of Theorem 1 remain satisfied under BIG sampling from $\mathcal{B}^*$. A motif k is observed in the sample $\mathcal{B}_s^*$, iff s_0 contains at least one of the nodes in β_k^*, regardless of the nodes in $\beta_k \setminus \beta_k^*$. Condition (iii) of Theorem 1 is satisfied provided the knowledge of β_k^*, given which the inclusion probabilities can be calculated by (2) and (3) on replacing β_k and β_l by β_k^* and β_l^*, respectively.

To ensure that BIG sampling from $\mathcal{B}^*$ is a feasible representation of T-SBS from G, we need to define β_k^* appropriately for the observed eligible motifs. By Corollary 1, BIGS representation is feasible for any motif consisting of connected nodes. Let the *observation diameter* of a motif $k \in \Omega(G)$ be

$$\phi_k = \max_{i \in M_k} d_{i,k}$$

which is finite for any motif of connected nodes with $|M_k| < \infty$. Then, by definition, an observed motif with finite ϕ_k is eligible for estimation under ϕ_k-SBS from G, provided we restrict its ancestors to $\beta_k^* = M_k$. The result below follows.

Theorem 2. *Provided finite observation diameter ϕ_k of all $k \in \Omega$, BIG sampling from $\mathcal{B}^*$ is a feasible representation for T-SBS from G, where $\beta_k^* = M_k$ and $T = \max_{k \in \Omega} \phi_k$.*

Additional sampling is not needed based on BIGS from $\mathcal{B}^*$ with restricted ancestors as a feasible representation of T-SBS from G. But fewer observed motifs are used compared to BIGS representation with $\mathcal{B}$, which would generally require additional sampling. So there

is a trade-off between statistical efficiency and operational cost. In case the uncertainty is too large to be acceptable, based on the eligible motifs in $\mathcal{B}_s^*$ under T-SBS with $T = \max_{k\in\Omega}\phi_k$, additional SBS may be administered. This raises the need to update the BIGS representation for T'-SBS, where $T' > T$.

Let $\beta^t(M_k)$ contain all the nodes outside of M_k, which have maximum geodesic distance t to M_k. That is, starting from any node in $\beta^t(M_k)$, it takes at most t stages of SBS to observe at least one of the nodes M_k. Under SBS beyond $T = \max_{k\in\Omega}\phi_k$, the nodes in $\beta^t(M_k)$ may be identified as ancestors of eligible motifs, for $t = 1, 2, ...$ Let the *diameter* of motif k be given by

$$\lambda_k = \max_{i,j\in M_k} \nu_{ij}$$

By Lemma 1, we have $\phi_k \leq 1 + \lambda_k$ given finite ϕ_k. The result below follows.

Theorem 3. *Provided finite observation diameter ϕ_k of all $k \in \Omega$, BIG sampling from $\mathcal{B}^*$ is a feasible representation for T-SBS from G, where $\beta_k^* = M_k \cup \beta^t(M_k)$ with $t \geq 1$, and $T = \max_{k\in\Omega} T_k$ with $T_k = \lambda_k + 2t$ for each $k \in \Omega$.*

6. Estimation under BIG sampling

For each motif $k \in \Omega(G)$, let y_k be an associated value, which is considered as an unknown constant. Let the target of estimation be the total of y_k over Ω, denoted by

$$\theta = \sum_{k\in\Omega} y_k \ .$$

In the case of $y_k \equiv 1$, θ is simply the total number of motifs in Ω, which is called a *graph total* (Zhang and Patone, 2017); more generally, θ is a total over Ω in a valued graph.

The two unbiased estimators of Birnbaum and Sirken (1965) can be applied to any graph sampling from G, provided a feasible BIGS representation of it satisfying conditions (i) - (iii) of Theorem 1. For simplicity below, we always denote the population BIG by $\mathcal{B}$, without distinguishing in notation whether restricted ancestors $\mathcal{B}^*$ are used for the eligible motifs. The HT estimator based on $\Omega_s = \Omega(\mathcal{B}_s)$ is given by

$$\hat{\theta}_y = \sum_{k\in\Omega_s} y_k/\pi_{(k)} = \sum_{k\in\Omega} \delta_k y_k/\pi_{(k)} \ , \tag{4}$$

where $\delta_k = 1$ if $k \in \Omega_s$ and 0 otherwise, and $\pi_{(k)}$ is given by (2), for any $k \in \Omega_s$. Generally, to calculate the inclusion probabilities $\pi_{(k)}$ and $\pi_{(kl)}$, we need to know β_k for each $k \in \Omega_s$. In the special case of SRS of s_0, we only need the cardinality of β_k to calculate $\pi_{(k)}$.

The HH-type estimator based on the initial sample s_0 is given by

$$\hat{\theta}_z = \sum_{i\in s_0} z_i/\pi_i = \sum_{i\in F} \delta_i z_i/\pi_i \quad \text{and} \quad z_i = \sum_{k\in\alpha_i} \omega_{ik} y_k \quad \text{and} \quad \sum_{i\in\beta_k} \omega_{ik} = 1 \ , \tag{5}$$

where $\delta_i = 1$ if $i \in s_0$ and 0 otherwise, and π_i is the inclusion probability of $i \in s_0$ under $p(s_0)$, and the ω_{ik}'s are constants of sampling, by which $\{y_k : k \in \Omega\}$ are transformed to the constructed measures $\{z_i : i \in F\}$. We let $\omega_{ik} = 0$ if $i \notin \beta_k$ or $k \notin \alpha_i$ in $\mathcal{B}$. As noted by Birnbaum and Sirken (1965), the estimator (5) is unbiased for θ since

$$\theta = \sum_{k\in\Omega} y_k = \sum_{k\in\Omega} y_k \big(\sum_{i\in\beta_k} \omega_{ik} \big) = \sum_{i\in F} \big(\sum_{k\in\alpha_i} \omega_{ik} y_k \big) = \sum_{i\in F} z_i \ .$$

Notice that in the special case of $|\beta_k| = 1$ for all $k \in \Omega$, there exits only one-one or one-many relationship between the sampling units in F and the motifs in Ω, just like when the $|M_k|$ elements are clustered in the sampling unit i under cluster sampling. The two estimators $\hat{\theta}_y$ and $\hat{\theta}_y$ are then identical. More generally, different choices of ω_{ik}'s would give rise to different estimates, such that $\hat{\theta}_z$ by (5) defines in fact a family of unbiased estimators. Birnbaum and Sirken (1965) consider the *equal-share* weights $\omega_{ik} = |\beta_k|^{-1}$. Under BIG sampling, this estimator and the HT-estimator have the same ancestral observation requirement. Patone (2020) proposes unequal weights $\omega_{ik} \propto |\alpha_i|^{-1}$. Additional sampling is generally needed to calculate these weights. For the feasible BIGS representation in Theorem 3, one may need upto $\phi_k + t$ extra stages to observe α_i for any $i \in \beta^t(M_k)$.

Theorem 4 below is a general result regarding the relative efficiency between $\hat{\theta}_y$ by (4) and $\hat{\theta}_z$ by (5), which applies to all situations where BIG sampling from $\mathcal{B}$ provides a feasible representation of the original graph sampling from G.

Theorem 4. *For $\hat{\theta}_y$ by* (4) *and $\hat{\theta}_z$ by* (5) *under BIG sampling from $\mathcal{B}$, we have*

$$V(\hat{\theta}_z) - V(\hat{\theta}_y) = \sum_{k\in\Omega}\sum_{l\in\Omega} \Delta_{kl} y_k y_l \quad \textit{where} \quad \Delta_{kl} = \sum_{i\in\beta_k}\sum_{j\in\beta_l} \frac{\pi_{ij}}{\pi_i \pi_j}\omega_{ik}\omega_{jl} - \frac{\pi_{(kl)}}{\pi_{(k)}\pi_{(l)}} \ .$$

7. Numerical work

Figure 1 shows a population graph G of 40 nodes and 72 edges. Let the motifs of interest be connected components of order $|M_k| \leq 4$, including node ($\mathcal{K}_1$), 2-clique (dyad, $\mathcal{K}_2$), 2-star ($\mathcal{S}_2$), 3-clique (triangle, $\mathcal{K}_3$), 4-clique ($\mathcal{K}_4$), 4-cycle ($\mathcal{C}_4$), 3-star ($\mathcal{S}_3$) and 3-path ($\mathcal{P}_3$). The 40 nodes are all known. The totals of the other motifs (illustrated in Figure 2) are

$$(\theta_{\mathcal{K}_2}, \theta_{\mathcal{S}_2}, \theta_{\mathcal{K}_3}, \theta_{\mathcal{K}_4}, \theta_{\mathcal{C}_4}, \theta_{\mathcal{S}_3}, \theta_{\mathcal{P}_3}) = (179, 72, 19, 3, 7, 141, 408) \ .$$

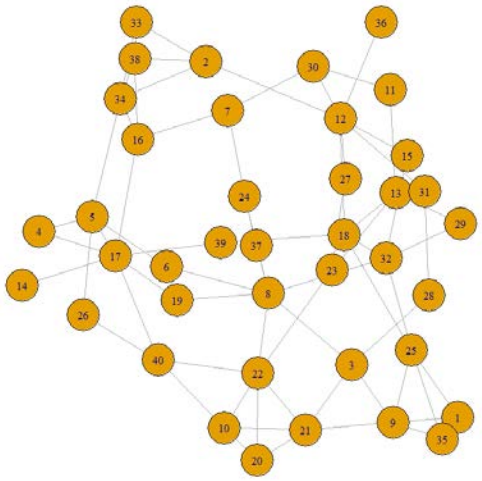

Figure 1: A population graph with $|U| = 40$ and $|A| = 72$.

For feasible BIGS representation with restricted ancestors $\beta_k^* = M_k$, the number of SBS stages required for the HT-estimator $\hat{\theta}_y$ by (4) and the HH-type estimator $\hat{\theta}_{z\beta}$ by (5) with weights $\omega_{ik} = |\beta_k|^{-1}$ is given by Theorem 2, i.e. $T = \phi_k$, whereas one may need up to ϕ_k additional stages for the estimator $\hat{\theta}_{z\alpha}$ using weights $\omega_{ik} \propto |\alpha_i|^{-1}$. Moreover,

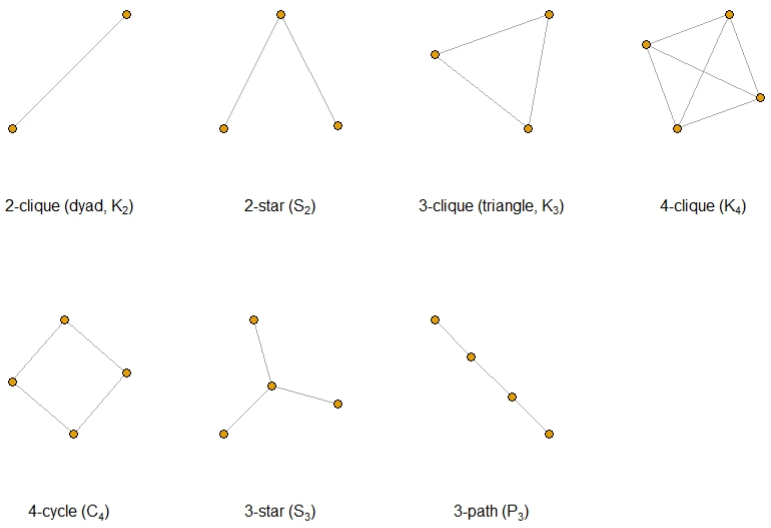

Figure 2: Motifs of interest

for BIGS representation with restricted ancestors $\beta_k^* = M_k \cup \beta(M_k)$, the number of stages required for $\hat{\theta}_y$ and $\hat{\theta}_{z\beta}$ is given by Theorem 3, i.e. $T = \lambda_k + 2t$ and $t = 1$, whereas up to $\phi_k + t = \phi_k + 1$ additional SBS stages may be needed for $\hat{\theta}_{z\alpha}$. Similarly in the case of $\beta_k^* = M_k \cup \beta^2(M_k)$ with $t = 2$.

Consider SBS of maximum 4 stages following SRS of s_0 with $|s_0| = 2$. Since the diameter of the population graph G is six here, a large part of it may already have been observed by 4-SBS; indeed, G is fully observed from 215 out of 780 possible initial samples. In addition, we consider induced OP following SRS of s, for which $s_{ref} = s \times s$. The size of s is set to be the expected number of observed nodes under $T = 1$ and $T = 2$, which are 9 and 21, respectively. Denote by $\hat{\theta}$ the resulting HT-estimator.

Table 1: Mean squared errors of graph total estimators under induced OP from SRS of size $n = 9$ or 21, and SBS of maximum 4 stages from SRS of initial sample of size 2.

	Estimator	$\mathcal{K}_2$	$\mathcal{S}_2$	$\mathcal{K}_3$	$\mathcal{K}_4$	$\mathcal{C}_4$	$\mathcal{S}_3$	$\mathcal{P}_3$
Induced OP, $n = 9$	$\hat{\theta}$	1 263	47 134	2 869	2 167	5 168	231 805	797 578
Induced OP, $n = 21$	$\hat{\theta}$	152	4 533	198	41	116	11 523	52 488
$\beta_k^* = M_k$	$\hat{\theta}_y$	471	5 269	193	10	38	5 092	27 717
	$\hat{\theta}_{z\beta}$	475	5 447	199	10	39	5 368	29 441
	$\hat{\theta}_{z\alpha}$	116	613	160	10	28	–	–
$\beta_k^* = M_k \cup \beta(M_k)$	$\hat{\theta}_y$	306	1 614	92	4	7	1 382	–
	$\hat{\theta}_{z\beta}$	281	1 485	98	5	7	1 403	–

In Table 1, we present the mean squared errors (MSEs) of the different estimators. Feasible BIGS representation is used for estimation under T-SBS. In case an estimator is not feasible for a certain motif using maximum 4-SBS, the result will be unavailable in the table. Induced OP is understandably much less efficient than incident OP, as the order of the motif of interest increases; compare e.g. the results for SRS of size 21 and 2-SBS, where both have the same expected number of nodes in the sample graph.

Under T-SBS from the population graph in Figure 1, the HT-estimator $\hat{\theta}_y$ and the HH-type estimator $\hat{\theta}_{z\beta}$ are about equally efficient for the motifs considered here. The HH-type estimator $\hat{\theta}_{z\alpha}$ can be much more efficient, especially for the lower-order motifs $\mathcal{K}_2$ and $\mathcal{S}_2$. Under SRS of s_0, the variance of the HH-type estimator (5) is minimised, if the constructed z_i's happen to be constant across the sampling units. With unequal-share weights, z_i is proportional to $|\alpha_i|$. Setting $\omega_{ik} \propto |\alpha_i|^{-1}$ tends to even out the z_i's, since a sampling unit with many successors will receive relatively little share from each motif observed from it, although its z-value is based on more motifs than another sampling unit with fewer successors. We refer to Patone (2020) for more discussions of $\hat{\theta}_{z\alpha}$.

8. Conclusion

Graph sampling (Zhang and Patone, 2017) provides a general statistical approach to study real graphs. We develop feasible BIGS representation that is applicable to a large number of graph sampling situations, which are based on different incident observation procedures. It avoids the recursive computations that are needed to calculate the inclusion probabilities of the sample motifs under T-stage snowball sampling (Zhang and Patone, 2017). It enables one to identify the motifs that are eligible to estimation in a given sample. It allows one to extend the scope of HH-type estimators developed for indirect sampling (Birnbaum and Sirken, 1965), providing a unified framework for achieving efficiency gains beyond the standard HT-estimator. A *generalised incidence weighting estimator* under BIGS is developed by Patone and Zhang (2020).

REFERENCES

Birnbaum, Z., and Sirken, M. (1965), "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates,", Vital and Health Statistics, PHS Publication No. 1000-Series 2, No. 11. National Center for Health Statistics, Washington, D. C.: U. S. Government Printing Office.

Frank, O. (1971), *Statistical Inference in Graphs*, Stockholm: FOA Repro.

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

Frank, O. (1980*a*), "Estimation of the number of vertices of different degrees in a graph," *Journal of Statistical Planning and Inference*, 4, 45–50.

Frank, O. (1980*b*), "Sampling and inference in a population graph," *International Statistical Review*, 48, 33–41.

Frank, O. (2011), "Survey sampling in networks,", The SAGE Handbook of Social Network Analysis. pages 389-403.

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

Goodman, L. (1961), "Snowball Sampling," *The Annals of Mathematical Statistics*, 32, 148–170.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, 14(4), pp. 333–362.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Lavallée, P. (2007), *Indirect Sampling*, New York, USA: Springer Science and Business Media.

Patone, M. (2020), "Topics of Statistical Analysis with Social Media Data,", Unpublished PhD thesis. University of Southampton, Southampton, United Kingdom.

Patone, M., and Zhang, L. (2020), "Incidence weighting estimation under bipartite incidence graph sampling,", arXiv:2004.04257[math.ST].

Sirken, M. (2005), "Network Sampling,", *Encyclopedia of Biostatistics*. John Wiley and Sons, Ltd., Online, DOI: 10.1002/0470011815.b2a16043.

Thompson, S. (1990), "Adaptive Cluster Sampling," *Journal of the American Statistical Association*, 85, 1050–1059.

Thompson, S. (1991), "Adaptive Cluster Sampling: Designs with Primary and Secondary Units," *Biometrics*, 47, 1103–1115.

Thompson, S. (2012), *Sampling (3rd ed.)*, New York: Wiley.

Zhang, L.-C., and Patone, M. (2017), "Graph Sampling," *Metron*, 75, 277–299.