The Survey Statistician, 2021, Vol. 83, 27–37.

New and Emerging Methods

Graph sampling: An introduction

Li-Chun Zhang[1,2,3]

[1]University of Southampton, UK, L.Zhang@soton.ac.uk
[2]Statistics Norway, Norway
[3]University of Oslo, Norway

Abstract

Representing a collection of relevant units by a graph allows one to incorporate the connections (or links) among the units in addition to the units themselves. One may be interested in the structure of the connections, or the links may provide effectively access to those units that are the primary interest. Either way, graph sampling provides a statistical approach to study real graphs. Just like sampling from finite populations, it is based on exploring the variation over all possible *sample graphs*, which can be taken from the given *population graph* according to a specified method of probability sampling, and design-based inference using a suitable graph sampling strategy is valid "whatever the unknown properties" (Neyman, 1934) of the population graph.

Keywords: graph, probability sampling, observation procedure, motif of interest, sampling strategy

1 Sampling from real graphs

Birnbaum and Sirken (1965) consider 'indirect sampling' of patients via an initial sample of medical centres. Since any in-scope patient may receive treatment at multiple places, not all of which are among the actual sample of medical centres, additional knowledge of all the out-of-sample treatment places of each sampled patient needs to be collected in order to calculate the patient's sample inclusion probability. In addition to the Horvitz-Thompson estimator (HTE, Horvitz and Thompson, 1952), Birnbaum and Sirken (1965) propose 2 unbiased estimators in this unusual situation.

Zhang (2020b) considers sampling to estimate the prevalence of an epidemic in a given population U, where one would like to increase the sample yield of *cases*, i.e. persons with $y_i = 1$ in contrast to *noncases* with $y_i = 0$, in order to improve the design efficiency. Let s_0 be an initial sample from U, with inclusion probability $\pi_i = \Pr(i \in s_0)$. Since the virus is transmitted via personal contacts, consider *adaptive network tracing*, where all the contacts of each case i in s_0 are included, and the procedure is repeated for them, and so on until no more cases can be added in this way. Let $\pi_{(i)} = \Pr(i \in s)$, where s is the final sample. Since any case i that is in contact with other cases can be included in s

by adaptive network tracing, even when it is not selected in s_0 initially, we achieve $\pi_{(i)} \geq \pi_i$. This is a special case of adaptive cluster sampling (ACS, Thompson, 1990) with binary y_i.

Zhang and Oguz-Alper (2020) develop the theory, which enables one to represent both the situations above as sampling from a *bipartite incidence graph (BIG)*. Patone and Zhang (2020) develop generally the *incidence weighting estimator (IWE)* under BIG sampling (BIGS), which encompasses all the estimators considered by Birnbaum and Sirken (1965). BIGS and the associated IWE form a flexible *graph sampling strategy*, which extends the finite-population (FP) sampling strategy consisting of a probability sampling design and the associated HTE. The BIGS-IWE strategy is applicable to many unconventional probability sampling techniques, which "are not explicitly stated as graph problems but which can be given such formulations" (Frank, 1977), including indirect sampling (Birnbaum and Sirken 1965; Lavalleè, 2007), network sampling (Sirken, 1970; 2005), adaptive cluster sampling (Thompson, 1990, 1991) and line-intercept sampling (Becker, 1992; Thompson, 2012). See Zhang and Oguz-Alper (2020) and Patone and Zhang (2020) for the relevant discussions.

As Zhang and Patone (2017) point out, in all the aforementioned situations, one is interested — rather conventionally — in some finite population total (or mean), where the connections (or links) among the relevant population elements and sampling units — more or less unconventionally — provide the access to the target population, which otherwise would have been ineffective or impractical to sample. Meanwhile, in sampling from arbitrary graphs generally, one is typically interested in the structure of the links themselves, often expressed in terms of a particular *motif*, which may simply be defined as a subgraph of specific characteristics. An early example is snowball sampling by Goodman (1961), where the motif of interest is 'pair with mutual relationships' in a special graph where all the nodes have out-degree one. In a series of work spanning over several decades (e.g. Frank, 1971, 1977, 1978, 1979, 1980, 1981, 2011), Ove Frank studies from this perspective graph sampling of motifs defined for nodes, dyads, triads (star, triangle), components, etc. Zhang and Patone (2017) provide a structure of *graph totals* of various motifs, to reflect the extended scope of investigation.

Thus taken together, representing a population of relevant units by a graph allows one to incorporate the connections (or links) among the units in addition to the units themselves. One may be either interested in the characteristics of the graph, or the links may provide effectively access to those units that are the primary interest. Either way, graph sampling provides a statistical approach to study real graphs. Just like sampling from finite populations, it is based on exploring the variation over all possible *sample graphs*, which can be taken from the given *population graph* according to a specified method of probability sampling, and design-based inference using a suitable graph sampling strategy is valid "whatever the unknown properties" (Neyman, 1934) of the population graph.

As much as graph sampling is versatile, it can be intricate when it comes to the formulation of graph sampling strategy in various situations. Below are three key elements in any case.

I. Definition of sample graph. Zhang and Patone (2017) define sample graph, where the specified sample observation procedure makes use of incident edges. Other observation procedures are conceivable which, in particular, may involve random jumps or teleporting to non-adjacent nodes. Tweaks of the definition of sample graph are needed accordingly.

II. Basis of inference. Zhang and Patone (2017) synthesise the existing graph sampling theory, where inference is based on the sample graph inclusion probabilities of the motif of interest. The IWE makes more extensive use of the same basis of inference, allowing for many unbiased estimators in addition to the HTE. More generally, inference can be based on other avaiable *sampling probabilities* associated with the given graph sampling method, as e.g. will be discussed for random walk sampling, which call for principally different strategies.

III. Eligible sample motifs. A motif that is observed in the sample graph is nevertheless 'ineligible' for estimation, if the required probabilities for inference cannot be calculated. Eligibility of a particular sample motif depends on the availability of the knowledge of its *ancestry* (Zhang and Patone, 2017). Essentially, apart form the actual way by which a motif is sampled, one needs to know how else it could have been sampled under the given sampling method. The concept of ancestry under graph sampling generalises the concept of *multiplicity* defined by Birnbaum and Sirken (1965), where it amounts to the knowledge of the out-of-sample medical centres for each sampled patient. Identification of eligible sample motifs is the key to any feasible graph sampling strategy.

In the rest of the paper, examples will be given to elaborate the points above. For the details that may be necessary for a fuller comprehension the reader is kindly referred to the relevant sources.

2 BIGS-IWE generalises FP-sampling and HT-estimation

Denote by $\mathcal{B} = (F, \Omega; H)$ a population BIG, where the node set is bipartitioned into F and Ω, such that (directed) edges exist only from F to Ω, denoted by $(i\kappa) \in H$, iff the selection of $i \in s_0 \subset F$ leads to that of κ from Ω. As explained and illustrated below, the strategy BIGS-IWE generalises the familiar strategy of 'FP-sampling and HT-estimation'.

Denote by $U = \{1, ..., N\}$ a *population* of size N. Let y_k be a constant associated with each $k \in U$, with population total $\theta = \sum_{k \in U} y_k$. Denote by s a sample from U, according to a method of probability sampling, where the sample inclusion probability $\Pr(k \in s)$ is either known in advance or can be calculated for the sample units afterwards. The HTE of θ is $\hat{\theta}_{HT} = \sum_{k \in s} y_k / \Pr(k \in s)$.

For element sampling, let $F = \Omega = U$, where $(i\kappa) \in H$ iff i and κ refer to the same population element. The correspond BIGS representation is given to the left in Figure 1. For cluster sampling, illustrated to the right in Figure 1, let F consist of the clusters (of which there are M in total) and $\Omega = U$ the elements that are nested in the clusters, where $(i\kappa) \in H$ iff element κ belongs to cluster i. Clearly, the strategy BIGS-HTE suffices for these familiar FP-sampling situations.

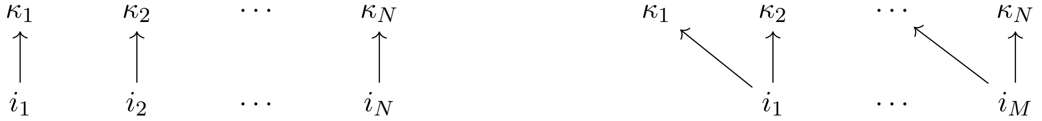

Figure 1: BIGS representation of finite-population sampling of elements (left) or clusters (right).

For indirect sampling of Birnbaum and Sirken (1965), let F consist of the medical centres and Ω the patients of interest, where $(i\kappa) \in H$ iff patient κ receives treatment at centre i. The BIG is illustrated in Figure 2, where the mapping from F to Ω can be many-many, instead of simply one-one or one-many as in Figure 1. Let us consider the elements I - III, in order to arrive at the strategy BIGS-IWE.

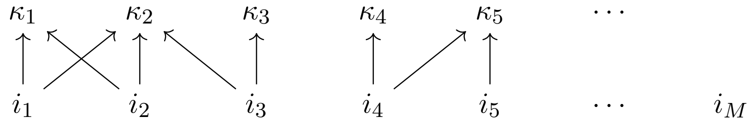

Figure 2: BIGS in general

I. Following the definition of Zhang and Patone (2017), the sample graph under BIGS from $\mathcal{B}$ is generally given as follows. Let s_0 be an initial sample from F, with sample inclusion probabilities π_i, π_{ij}, etc. Apply the *incident forward observation procedure*, by which all the out-edges from s_0 are included in the sample of edges, denoted by $H_s = \{(i\kappa) : (i\kappa) \in H, i \in s_0\}$. The sample nodes are the union of s_0 and those incident to the edges in H_s, i.e. $s_0 \cup \Omega_s$, where $\Omega_s = \alpha(s_0) = \cup_{i \in s_0} \alpha_i$, and

$\alpha_i = \{\kappa : (i\kappa) \in H\}$ are the *successors* of i in $\mathcal{B}$. The sample graph under BIGS is given by

$$\mathcal{B}_s = (s_0, \Omega_s; H_s).$$

II. The inference is based on the sample inclusion probabilities. The fact that each κ in Ω_s can possibly be accessed via multiple sampling units in F calls for the concept of ancestry under graph sampling. For instance, suppose $i_1 \in s_0$ but not i_2 or i_3. To calculate the inclusion probability of the patient $\kappa_2 \in \alpha_{i_1}$, one must collect the information that it receives treatment at i_2 and i_3 as well, i.e. how else it could have been sampled under BIGS here.

III. The ancestry knowledge of $\kappa \in \Omega_s$ is secured and it is eligible for estimation of $\theta = \sum_{\kappa \in \Omega} y_\kappa$, where y_κ is a constant associated with each $\kappa \in \Omega$, provided the observation of $\beta_\kappa \setminus s_0$, where $\beta_\kappa = \{i \in F : (i\kappa) \in H\}$ are the *predecessors* of κ in $\mathcal{B}$, although $\beta_\kappa \setminus s_0$ are not part of the sample graph $\mathcal{B}_s$. It follows that all the nodes in Ω_s are eligible, provided the observation of $\beta\big(\alpha(s_0)\big) \setminus s_0$ in addition to $\mathcal{B}_s$, where $\beta\big(\alpha(s_0)\big) = \cup_{\kappa \in \alpha(s_0)} \beta_\kappa$.

Let $W_{i\kappa}$ be the *incidence weight* associated with each edge $(i\kappa) \in H_s$. The IWE of θ is given by

$$\hat{\theta} = \sum_{(i\kappa) \in H_s} W_{i\kappa} \frac{y_\kappa}{\pi_i} \tag{1}$$

(Patone and Zhang, 2020), where $\pi_i = \Pr(i \in s_0)$ is also the probability that $(i\kappa)$ is included in $\mathcal{B}_s$ under BIGS. While the HTE is defined for Ω_s, the IWE is defined for the sample BIG edge set $H(\mathcal{B}_s) = H_s$, where each $(i\kappa)$ in H_s is incident to i in s_0 and κ in Ω_s. Patone and Zhang (2020) show that the IWE encompasses all the estimators considered by Birnbaum and Sirken (1965). In particular, the HTE is a special case, where $W_{i\kappa}$ varies according to $s_\kappa = s_0 \cap \beta_\kappa$, subjected to the condition that ensures unbiased estimation of θ over repeated sampling: for any $\kappa \in \Omega$,

$$\sum_{i \in \beta_\kappa} E(W_{i\kappa} | i \in s_0) = 1.$$

This generalises the result of Birnbaum and Sirken (1965) for constant weights, denoted by $\omega_{i\kappa}$ for distinction, which requires $\sum_{i \in \beta_\kappa} \omega_{i\kappa} = 1$ for any $\kappa \in \Omega$, including $\omega_{i\kappa} = 1/m_\kappa$ and $m_\kappa = |\beta_\kappa|$.

3 BIGS-IWE for unconventional sampling: ACS as an example

3.1 ACS with binary outcome variable

Let U be the population of size N and $\mu = \theta/N$ the prevalence of interest. Adaptive network tracing requires the population U to be represented by the population graph $G = (U, A)$ where, in addition to the node set U, the edge set A contains all the relevant contacts. We shall treat the graph as undirected and simple, where $(ij), (ji) \in A$ if persons i and j are in-contact, and there is only one edge either way regardless of the frequency or intensity of the contact between i and j.

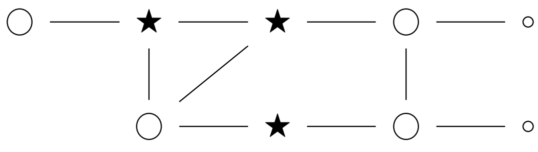

Figure 3: Cases ★, noncase edge nodes ◯, other noncases ∘

Figure 3 illustrates a part of such a population graph G, with stars for cases and circles for noncases. In particular, all the cases are partitioned into *case networks*, where those in the same network all have $y_i = 1$ and are connected to each other in G; two case networks are shown in Figure 3. Next, a (network) *edge node* is a noncase that is adjacent to at least one case network; the four edges nodes in Figure 3 are shown as bigger circles than the other noncases.

ACS from G employs contact tracing starting from an initial sample s_0 from U, which is adaptive because tracing is only applied to the contacts of ★ but not ◯ or ∘. The final sample s by ACS can be divided into three parts: (i) a set of case networks, (ii) the edge nodes, and (iii) the remaining noncase nodes in the initial sample s_0 which do not belong to (i) and (ii). Zhang (2020b) considers the efficacy of several ACS designs for cross-sectional as well as change estimation of prevalence. These graph sampling methods allow one to unite tracing for combating the disease *and* sampling for estimating the prevalence during an epidemic outbreak.

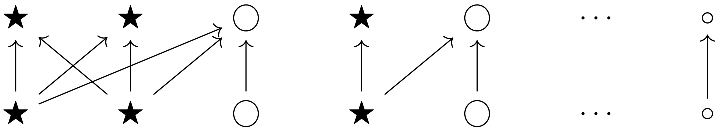

Figure 4: BIGS representation of ACS from G with binary outcome variable

Following Zhang and Oguz-Alper (2020), one can apply the BIGS-IWE strategy to ACS from G above, thereby allow other unbiased estimators in addition to the HTE. Let $F = \Omega = U$ in $\mathcal{B}$, and let H contain the incident observation relationships among the nodes in G under adaptive network tracing. For instance, let the two leftmost ★ in Figure 4 be the two in the same network in Figure 3, and let ◯ next to them in Figure 4 be one of their edge nodes in Figure 3. Under ACS from G, the selection of either ★ in s_0 leads to all the three of them to be included in the sample s, yielding the corresponding edges from these two ★ in $\mathcal{B}$. Meanwhile, selecting any ◯ in s_0 does not lead on to any adjacent case network, such that ◯ has only an edge to itself in $\mathcal{B}$. Similarly, the other two edge nodes in Figure 3 can be included in $\mathcal{B}$, which are omitted here to avoid cluttering the figure visually.

Thus, the sample graph under BIGS from $\mathcal{B}$ is the same as that by ACS from G. The inference basis is still the relevant sample inclusion probabilities. Since each sample case is observed together with its network under ACS, the knowledge of its ancestry is secured for the BIGS representation with $\mathcal{B}$ defined above. But ancestry is generally unclear for any edge node ◯ in Figure 4, since we would not observe any of its ★-ancestors in a case network unless that network happens to intersect s_0. *However, this does not matter here, since a noncase κ in Ω with $y_\kappa \equiv 0$ contributes nothing to the IWE* (1) *regardless its inclusion probability.* Thus, using only the sample case nodes as the eligible motifs in Ω_s, the BIGS-IWE is a feasible strategy for ACS from G with binary outcome variable.

3.2 ACS with continuous outcome variable

Simply ignoring the edge nodes would not be valid for ACS with continuous outcome variable, where an edge node generally has a non-zero value below the threshold chosen for adaptive sampling. Thompson (1990) proposes an inferential approach, where one modifies two of the estimators of Birnbaum and Sirken (1965). Zhang and Oguz-Alper (2020) develop the BIGS-IWE strategy. Let us illustrate their approach here using the example of Thompson (1990).

The population U consists of $N = 5$ spatial grids, with associated $y_U = \{1, 0, 2, 10, 1000\}$ for the amount of species of interest. Each grid has either one or two neighbours which are adjacent in the undirected graph $G = (U, A)$ below, where we simply denote each grid (or node) by its y-value as Thompson (1990). This is a *valued* graph where G is known but the associated y_U are unknown.

$$G: \qquad 1 \text{ —— } 0 \text{ —— } 2 \text{ —— } 10 \text{ —— } 1000$$

Figure 5: Graph for ACS (Thompson, 1990)

Given an initial sample s_0 of size 2 by simple random sampling (SRS) from U, one would survey all the adjacent grids (in both directions if possible) of a sample grid i if y_i exceeds the threshold value 5 but not otherwise, and so on. A network in G may consist of one or more connected nodes all with y-values above the threshold such as $\{10, 1000\}$ here, or it may be a single node with y-value up to the threshold, some of which are edge nodes such as node 2 here. The interest is to estimate the mean amount of species per node, denoted by $\mu = \theta/N$, where $\theta = \sum_{i \in U} y_i$.

Since the sample inclusion probability of any edge node is generally unknown under ACS from G, Thompson proposes to modify the HTE, such that an edge node i is used for estimation (i.e. eligible) only when $i \in s_0$ directly, the probability of which is $\pi_i = \Pr(i \in s_0) = n/N$ under SRS of s_0. Similar modification can be applied to the 2nd estimator of Brinbaum and Sirken (1965), which is referred to as the *Hansen-Hurvitz (HH) type* estimator by Thompson (1990).

Zhang and Oguz-Alper (2020) denote the strategy of Thompson (1990) by $(\mathcal{B}, \hat{\theta}^*_{HT})$ when the modified HTE is used as the estimator, where the population $\mathcal{B}$ has $F = \Omega = U$ and the edge set H contains all the observational links under ACS from G. They observe that it is as well possible to modify the sampling when constructing a feasible strategy, say, (ACS*, HT) or (ACS*, HH). In particular, they use BIGS as ACS*, in which case the IWE would unify and generalise the HTE and HH-type estimator.

For a generally feasible strategy with BIGS one can use instead $\mathcal{B}^* = (U, U; H^*)$ in Figure 6. The observational links $(10, 2)$ and $(1000, 2)$ under ACS from G are removed to ensure ancestral observation in $\mathcal{B}^*$. For instance, given $s_0 = \{0, 2\}$, the observation procedure of ACS means 10 and 1000 are not observed, as in $\mathcal{B}^*$ where 2 in $\Omega(\mathcal{B}^*)$ has only itself as the ancestor in $F(\mathcal{B}^*)$. One can now use the unmodified HTE under BIGS from $\mathcal{B}^*$, as a special case of IWE, denoted by $(\mathcal{B}^*, \hat{\theta}_y)$.

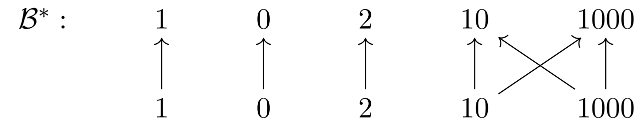

Figure 6: A feasible BIGS-IWE strategy for ACS

The two strategies $(\mathcal{B}, \hat{\theta}^*_{HT})$ and $(\mathcal{B}^*, \hat{\theta}_y)$ lead to the same estimator, since the eligible sample nodes in Ω_s are the same under both. The difference is that applying the Rao-Blackwell method does not change $\hat{\theta}_y$ under BIGS from $\mathcal{B}^*$, whereas it changes $\hat{\theta}^*_{HT}$ under BIGS from $\mathcal{B}$.

Another possible strategy using BIGS in this particular setting is to make an edge node ineligible, if itself is selected in s_0 but not its neighbouring above-threshold network, with $\mathcal{B}^\dagger$ in Figure 7. Denote this strategy by $(\mathcal{B}^\dagger, \hat{\theta}_y)$. It is feasible here because the egde node 2 has only one above-threshold neighbouring network in G, i.e. $\{10, 1000\}$; but it would be infeasible generally provided an edge node has two or more such networks in G.

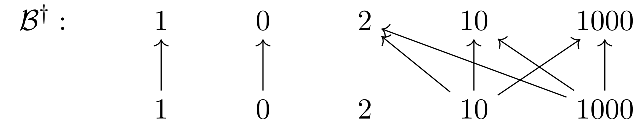

Figure 7: Another BIGS-IWE strategy for ACS

The BIGS-IWE strategy extends the inferential approach to ACS pioneered by Thompson (1990), where one can modify *either* part of a standard strategy (sampling, estimator) when it is otherwise infeasible in a given situation. For the example discussed above, Table 1 taken from Zhang and Oguz-Alper (2020) provides the numerical details of the three strategies $(\mathcal{B}, \hat{\theta}^*_{HT})$, $(\mathcal{B}^*, \hat{\theta}_y)$ and $(\mathcal{B}^\dagger, \hat{\theta}_y)$.

Table 1: Strategies using BIGS for ACS from $G:\ 1-0-2-10-1000$.

	$(\mathcal{B}, \hat{\theta}^*_{HT})$		$(\mathcal{B}^*, \hat{\theta}_y)$		$(\mathcal{B}^\dagger, \hat{\theta}_y)$	
s_0	Ω_s	$\hat{\mu}^*_{HT}$	Ω_s	$\hat{\mu}_y$	Ω_s	$\hat{\mu}_y$
1,0	1,0	0.500	1,0	0.500	1,0	0.500
1,2	1,2	1.500	1,2	1.500	1	0.500
0,2	0,2	1.000	0,2	1.000	0	0.000
1,10	1,10,*2*,1000	289.071	1,10,1000	289.071	1,10,2,1000	289.643
1,1000	1,1000,*2*,10	289.071	1,1000,10	289.071	1,1000,2,10	289.643
0,10	0,10,*2*,1000	288.571	0,10,1000	288.571	0,10,2,1000	289.143
0,1000	0,1000,*2*,10	288.571	0,1000,10	288.571	0,1000,2,10	289.143
2,10	2,10,1000	289.571	2,10,1000	289.571	2,10,1000	289.143
2,1000	2,1000,10	289.571	2,1000,10	289.571	2,1000,10	289.143
10,1000	10,1000,*2*	288.571	10,1000	288.571	10,1000,2	289.143
Variance		17418.4		17418.4		17533.7

4 Graph sampling: On to general theory

Sampling from arbitrary graph $G = (U, A)$ involves a number of conceptual generalisations of BIG sampling discussed above. Given limited space, we focus our discussion here on the following.

- So far we have only seen examples of motifs defined for the nodes of G. Zhang and Patone (2017) define generally motifs and their graph total, i.e. instead of population total over U.
- There are many observation procedures in a graph, which make use of the edges incident to an initial sample of nodes. Zhang and Patone (2017) discuss induced and incident procedures.
- Repeating an incident observation procedure leads to *multiwave* sampling. *T-wave snowball sampling* (T-SBS, Zhang and Patone, 2017) is the probabilistic version of breath-first search in graphs, and *targeted random walk (TRW)* is that of depth-first search.
- Zhang and Patone (2017) consider inference based on the graph sample inclusion probabilities. Depending on the observation procedure, other sampling probabilities may be necessary.

Definitions Let $G = (U, A)$ consist of nodes U and edges A. Let A_{ij} contain the edges from i to j, and $a_{ij} = |A_{ij}|$. Attaching values to U or A yields a *valued graph*. One may consider a graph to be the *structure* of a valued graph. We do not consider separately sampling from graphs or value graphs. Generally speaking, a graph sampling method may depend on the values associated with G, and the values associated with the sample graph G_s are observed together with G_s.

Let M be a subset of U. Let $G(M)$ be the subgraph *induced* by M, whose edge set is given by $\{A_{ij} : (i,j) \in M\}$. A subgraph $G(M)$ with specific characteristics is called a *motif*, denoted by $[M]$. For example, $[i : a_{i+} = 3]$ is a motif of node with out-degree 3, $[i,j : a_{ij}a_{ji} = 1]$ of a node pair with mutual simple relationship, and $[i,j : a_{ij} + a_{ji} = 0]$ of a non-adjacent node pair.

Let $y\big(G(M)\big)$, or simply $y(M)$, be a function of $G(M)$. Let Ω contain all the relevant M. Let

$$\theta = \sum_{M \in \Omega} y(M) \tag{2}$$

be the *graph total* over Ω. It is said to be the *k-th order*, if $|M| = k$ for all $M \in \Omega$. Although It is possible to let Ω in (2) be the set of motifs of interest directly, and let y_κ be a function of motif κ, it can be convenient if the summation is over all the relevant node sets. For instance, the motifs $[i,j : a_{ij}a_{ji} = 1]$ can be enumerated over $\Omega = \{(i,j) : i \neq j \in U\}$, with $y(i,j)$ as the corresponding counter. If Ω is the set of these motifs, then writing $\theta = |\Omega|$ is more natural than $\theta = \sum_{\kappa \in \Omega} 1$.

Zhang and Patone (2017) consider induced or incident *observation procedure* given an initial sample of nodes, denoted by s_0. Take $G: \ a \quad b \to c \to d$ for example. Let $s_0 = \{b, d\}$. We observe none of the edges if the observation procedure is induced, or the edge (bc) if it is incident forward, or (cd) if incident backward, or both (bc) and (cd) if incident reciprocal.

Ove Frank studies sampling of node sets or motifs using such observation procedures, where a sample of motifs from the population of motifs is conceived in analogy to a sample s from the population U. Zhang and Patone (2017) define the *sample graph* $G_s = (U_s, A_s)$ as a subgraph of G.

- Initial sample of nodes $s_0 \subset U$, with $p(s_0)$, π_i, π_{ij}, etc.
- Application of the specified observation procedure, starting from s_0.
- Specify the reference set $s_{\text{ref}} \subset U \times U$, such that $A_s = A \cap s_{\text{ref}}$
- Let $U_s = s_0 \cup \text{Inc}(A_s)$, where $\text{Inc}(A_s)$ denotes the nodes incident to the edges in A_s.

Compared to sampling from finite populations, a defining feature of sampling from graphs is that one uses the edges. The definition of sample graph above includes the situation, where the initial node sample s_0 is given as the nodes that are incident to a sample of edges directly selected from A. Direct sampling of edges may be useful e.g. if G is known but is too large to be counted. It is then possible for the observation procedure to specify that no additional edges need to be sampled.

It is convenient to specify the sample edges A_s via s_{ref}, which explicates the parts of the adjacency matrix $[a_{ij}]$ that are observed given s_0 and the observation procedure. Take again $G: \ a \quad b \to c \to d$ for example. Let the rows and columns of $[a_{ij}]$ be arranged in the order a, b, c, d. The set s_{ref} given $s_0 = \{b, d\}$ and the various observation procedures are shown in $\boxed{1/0}$ below.

$$\begin{array}{cccc}
\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boxed{0} \\ 0 & 0 & 0 & 1 \\ 0 & \boxed{0} & 0 & 0 \end{bmatrix} &
\begin{bmatrix} 0 & 0 & 0 & 0 \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & 0 & 0 & 1 \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix} &
\begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ 0 & \boxed{0} & 1 & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ 0 & \boxed{0} & 0 & \boxed{0} \end{bmatrix} &
\begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix} \\
\text{Induced} & \text{Incident forward} & \text{Incident backward} & \text{Incident reciprocal}
\end{array}$$

T-SBS To keep focus, let *incident OP* stand for incident forward observation procedure from now on. Note that in undirected graphs, incident is the same as incident reciprocal. Given an initial *seed* sample s_0 from U, let $s_1 = \alpha(s_0) \setminus s_0$ be the 1st-wave seed sample. Repeat the incident OP for s_1, which may or may not result in a non-empty 2nd-wave seed sample $s_2 = \alpha(s_1) \setminus (s_0 \cup s_1)$. Carry on this way yields the seed samples s_3, ..., s_T. The *seed sample of T-SBS* is given by $s = \bigcup_{r=0}^{T-1} s_r$.

The reference set s_{ref} of T-SBS is $s \times U$ in directed graphs or $s \times U \cup U \times s$ in undirected graphs. A motif $[M]$ is observed in the sample graph G_s iff $M \times M \in s_{\text{ref}}$. Frank and Snijders (1994) consider 1-SBS for node (1st-order) graph totals. Zhang and Patone (2017) develop the HTE for finite-order graph totals under T-SBS. The basis of inference is the graph sample inclusion probabilities.

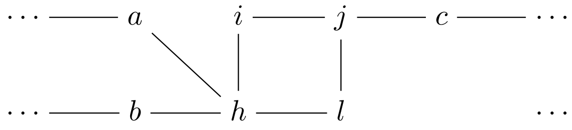

Figure 8: An example of 4-cycle $[h, i, j, l]$

However, not all the motifs observed in G_s are eligible for estimation. Consider the 4-cycle motif $[M]$ with $M = \{h, i, j, l\}$ in Figure 8. We need two waves to observe a 4-cycle, starting from any node in M, so that it is observed under 3-SBS starting from any of a, b, c. If only a is selected in s_0, then we would observe this 4-cycle under 3-SBS, as well as b as another of its ancestors, but not c, so that its sample inclusion probability under 3-SBS cannot be calculated and it remains ineligible.

Zhang and Oguz-Alper (2020) develop the theory for eligible sample motifs under T-SBS. One could carry on SBS further, until one has obtained all the ancestors of the sample motifs observed under T-SBS. One could use only the eligible sample motifs under T-SBS, in a manner resembling the modified HTE that excludes the edge notes under ACS. As explained earlier, it is also possible to modify the sampling to allow for the unmodified HTE. Zhang and Oguz-Alper (2020) develop feasible BIGS representation for T-SBS with given T. One can then apply the IWE instead of only the HTE using this BIGS-IWE strategy, tailored to the number of waves when sampling is terminated.

Targeted random walk One can envisage a discrete-time walk in a graph as travelling from one city to another via the existing roads that connect the cities. In a *random walk*, one takes randomly one of the possible roads out of the current city, repeat the same at the next city, and so on. A walk reaches gradually its *equilibrium*, if the chance that one visits a given city at a given time depends less and less on the particular starting point. The *stationary probability* that a walk takes one to a given city is the fraction of times the city is visited when the walk is at equilibrium.

Random walk in large and possibly dynamic graphs has been used in many disciplines (Masuda et al., 2017), including Google PageRank (Brin and Page, 1998), especially if the walk is fast-moving. For a connected undirected graph, the stationary probabilities of a random walk, denoted by π_i for $i \in U$, are known up to a proportional constant for the nodes visited, but not the ones yet to be visited. Thompson (2006a) applies the Metropolis-Hastings acceptance mechanism to the proposed moves, in order to achieve other targeted stationary probabilities, such as $\pi_i \propto$ degree+1. This requires one to observe *all* the neighbours of *all* the adjacent nodes of the current one, which may be impractical. Avrachenkov et al. (2010) devise an elegant random walk that requires only the knowledge of the adjacent nodes at each step time, for which the stationary probabilities at equilibrium is known up to a proportional constant for undirected graphs. The disconnected components are accommodated by random jumps via an imaginary node. We refer to it as *targeted random walk (TRW)*.

The random-walk inclusion probability of a node is intractable. Insofar as the stationary probability π_i is the same for a given node i at any time step for a random walk at equilibrium, and π_i is only known up to a proportional constant, approximately unbiased estimation is possible, e.g. for the ratio between two *node* totals using the generalised ratio estimator (Thompson, 2006a).

For other finite-order graph totals generally, Zhang (2020a) demonstrates that inference can be based on the *stationary successive sampling probability (S3P)* of any subsequence from the TRW states $\{X_0, X_1, ..., X_T\}$. For example, suppose $(X_t, X_{t+1}) = (i, j)$ where $a_{ij} = 1$. As long as both X_t and X_{t+1} belong to the seed sample s of the TRW, following the same definition of seed sample of T-SBS above, one can observe e.g. all the triangles (i, j, h) in the graph. The S3P of the actual sampling sequence $(X_t, X_{t+1}) = (i, j)$ is $\pi_i p_{ij}$, where p_{ij} is the corresponding transition probability. Moreover, all the possible (X_t, X_{t+1}) that lead to the observation of the same triangle (i, j, h) are called the *equivalent successive sampling sequences (ES3)*, including $(X_t, X_{t+1}) = (j, i), (h, i), (i, h), (h, j), (j, h)$ in addition to (i, j). The ES3 of a motif $[M]$ constitutes its multiplicity under TRW sampling sequence-by-sequence. A motif $[M]$ is eligible for estimation, if its ES3 is observed under the TRW. This yields a BIGS representation of TRW, with the motifs of interest in Ω and their ES3s in F. By this development, BIGS-IWE becomes a feasible strategy for any function of finte-order graph totals under TRW, as long as the function is invariant towards the unknown proportional constant in π_i.

5 Some topics for future research

Graph sampling is clearly the future of sampling. We have provided a brief introduction to it mainly by examples. The references contain many details that may be helpful for further reading.

A topic for future research is other possible bases of inference in various graph sampling situations. For instance, Thompson (2006b) considers adaptive web sampling where, at each wave, a subset of the already sampled nodes are used as the seeds for random walks from them. The wave-by-wave conditional sampling probabilities are used for estimation of node totals, together with all the seed sample selection probabilities. A theory is needed for other finite-order graph totals.

It is intriguing to consider other parameters than graph totals (2) and functions of them. Newman (2010) is an excellent source of candidates, many of which can be characterised as 'a local parameter dependent on the whole graph'. For example, the in-degree of a node i is a local parameter that only depends on its in-edges, but not the rest of the graph. The betweenness of a node provides an example of what we have in mind. As can be seen below, the shortest-path (SP) betweenness of ★ is 0, which is defined as the fraction of SPs between pairs of nodes in a graph passing through it, because it is always 'short-circuited' by the two ○ nodes. Whereas ★ has a high random-walk (RW) betweenness (Newman, 2005), which is defined as the fraction of RWs between pairs of nodes in a graph passing through it. How would a sampling strategy look like for such parameters?

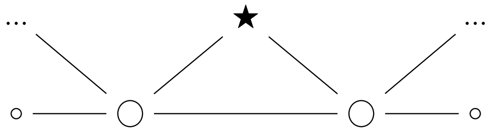

Figure 9: An example for betweenness

Graph sampling poses not least an enormous opportunity for computation. Efficient sample motif counting algorithms are obviously critical in applications, and their availability can influence the theory in return. For instance, Frank (1971) considers 'telepathy-like' observation of whether two nodes are connected, without explicating any path between them. One can envisage the possibility the relevant algorithm is so fast that it is virtually instant when the graph is known, e.g. depending on the data structure implemented. However, the graph may be so large or dynamic that sampling is still needed for 'graph compression'. The availability of such *remote* observation procedures could easily lead to other possibilities of sample graph, basis of inference and graph sampling strategy.

References

Avrachenkov, K., Ribeiro, B. and Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. *Research report*, RR-7394, INRIA. inria-00520350

Becker, E.F. (1991). A terrestrial furbearer estimator based on probability sampling. *The Journal of Wildlife Management*, 55:730–737.

Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of lRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30: 107–117.

Frank, O. (1971). *Statistical inference in graphs*. Stockholm: Försvarets forskningsanstalt.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235–264.

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

Frank, O. (1979). Sampling and estimation in large social networks. *Social networks*, 1(1):91–101.

Frank, O. (1980). Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique*, 48:33–41.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological methodology*, 12:110–155.

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*, 10:53–53.

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

Lavalleѐ, P. (2007). *Indirect Sampling*. Springer.

Masuda, N., Porter, M.A. and Lambiotte, R. (2017) Random walks and diffusion on networks. *Physics Reports*, 716-717: 1–58. `http://dx.doi.org/10.1016/j.physrep.2017.07.007`

Newman, M.E.J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27, 39–54.

Newman, M.E.J. (2010). *Networks: An Introduction*. Oxford University Press.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Patone, M. and Zhang, L.-C. (2020). Incidence weighting estimation under bipartite incidence graph sampling. `https://arxiv.org/abs/2004.04257`

Sirken, M.G. (2005). *Network Sampling*. In *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a16043

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65:257–266.

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

Thompson, S.K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, 47:1103–1115.

Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32, 11–24.

Thompson, S.K. (2006b). Adaptive Web Sampling. *Biometrics*, 62, 1224–1234.

Thompson, S.K. (2012). *Sampling*. John Wiley & Sons, Inc.

Zhang, L.-C. (2020a). Targeted random walk sampling from large dynamic graphs. *Talk presented at University of Perugia, November 3, 2020.*

Zhang, L.-C. (2020b). Sampling designs for epidemic prevalence estimation. `https://arxiv.org/abs/2011.08669`

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

Zhang, L.-C. and Oguz-Alper, M. (2020). Bipartite incidence graph sampling. `https://arxiv.org/abs/2003.09467`