Graph Sampling

Li-Chun Zhang

Contents

Symbols

$G = (U, A)$	population graph
U	nodes in G, population of units
A	edges in G
Ω	collection of networks/motifs in G, study units in $\mathcal{B}$
M	index of set of nodes
$i, j, h, ...$	index of node
$\kappa, \ell, ...$	index of motif
θ	graph total
μ	graph parameter
$G_s = (U_s, A_s)$	sample graph
U_s	sample of nodes in G_s
A_s	sample of edges in G_s
Ω_s	sample of networks/motifs, study units
s_{ref}	reference set, parts of adjacency matrix
s_0	initial node sample
s	seed sample
$\mathcal{B} = (F, \Omega; H)$	bipartite incidence graph (BIG)
F	collection of sampling units
H	edges in BIG
$\mathcal{B}_s = (s_0, \Omega_s; H_s)$	sample BIG
H_s	sample of edges in $\mathcal{B}_s$

Abbreviations

FPS	finite-population sampling
HT	Horvitz-Thompson
OP	observation procedure
SRS	simple random sampling
BIG	bipartite incidence graph
BIGS	bipartite incidence graph sampling, BIG sampling
IWE	incidence weighting estimator
HTE	Horvitz-Thompson estimator, HT-estimator
HH	Hansen-Hurwitz
RB	Rao-Blackwell
PIDA	probability and inverse-degree adjusted
RE	relative efficiency
LIS	line-intercept sampling
ACS	adaptive cluster sampling
SBS	snowball sampling
TSBS	T-wave snowball sampling, T-wave SBS
MH	Metropolis-Hastings
TWS	targeted walk sampling
TRW	targeted random walk
TRWS	targeted random walk sampling, TRW sampling
TTRWS	T-step targeted random walk sampling, T-step TRWS
S3P	stationary successive sampling probability
AS3	actual sampling sequence of states
ES3	equivalent sampling sequence of states

Chapter 1

Introduction to graph sampling

1.1 Sampling from finite populations

Denote by $U = \{1, ..., N\}$ a *population* of size N. Denote by s a *sample* from U, $s \subset U$, according to some specified method of sampling. Let $\pi_i = \Pr(i \in s)$ be the *sample inclusion probability* of $i \in U$. *Probability sampling* is the case if $\pi_i > 0$ for any $i \in U$, and π_i is either known before the sample is drawn or can be calculated afterwards.

Let y_i be an unknown value associated with each $i \in U$. Let $Y = \sum_{i \in U} y_i$ be its population total. The Horvitz-Thompson (HT) estimator of Y is given by

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in U} \delta_i \frac{y_i}{\pi_i}$$

where $\delta_i = \mathbb{I}(i \in s)$. It is unbiased over repeated sampling, $E(\hat{Y}_{HT}) = Y$, with variance

$$V(\hat{Y}_{HT}) = \sum_{i \in U} \sum_{j \in U} Cov(\delta_i, \delta_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \sum_{i \in U} \sum_{j \in U} (\frac{\pi_{ij}}{\pi_i \pi_j} - 1) y_i y_j := \sum_{i \in U} \sum_{j \in U} v_{ij}$$

where $\pi_{ij} = E(\delta_i \delta_j) = \Pr(i \in s, j \in s)$ is the *joint* sample inclusion probability of (i, j), or the second-order inclusion probability. Since $V(\hat{Y}_{HT})$ is the total of v_{ij} over $U \times U$, whose element (ij) has inclusion probability π_{ij} in $s \times s$, an estimator is given by

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{v_{ij}}{\pi_{ij}} = \sum_{i \in s} \sum_{j \in s} (\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}) y_i y_j$$

1.2 Graph, motif, graph parameter

Representing a finite population U by a graph allows one to incorporate the connections (or links) among the population units in addition to the units themselves.

1.2.1 Graph

A graph $G = (U, A)$ consists of a set of nodes U and a set of edges A, where $|U| = N$ and $|A| = R$ are the *order* and *size* of G, respectively. Attaching values to U or A yields a *valued graph.* A graph is the *structure* of the corresponding valued graph.

Let A_{ij} be the set of edges from i to j, such that $A = \bigcup_{i,j\in U} A_{ij}$. Let $a_{ij} = |A_{ij}|$. The graph is a *multigraph* if $a_{ij} > 1$ for some $i, j \in U$; it is a *simple* graph otherwise.

By default a graph is *directed*, or a *digraph.* The out-edges of node i are $A_{i+} = \bigcup_{j\in U} A_{ij}$ and its in-edges are $A_{+i} = \bigcup_{j\in U} A_{ji}$. The out-degree is $a_{i+} = |A_{i+}| = \sum_{j\in U} a_{ij}$ and the in-degree is $a_{+i} = |A_{+i}| = \sum_{j\in U} a_{ji}$. A graph is *undirected* if there is no distinction between in- and out-edges, $A_{ij} \equiv A_{ji}$, where $d_i = a_{i+} = a_{+i}$ is the degree of node i.

Two nodes i and j are *adjacent* if there exists at least one edge between them, in which case $a_{ij} + a_{ji} > 1$. For any edge in A_{ij}, i is its initial node and j its terminal node. An edge is *incident* to its initial and terminal nodes, and vice versa. Note that adjacency refers to relationship between nodes, as objects of the same kind, whereas incidence refers to relationship between nodes and edges, as objects of different kinds.

Let α_i be the *successors* of i, which are the terminal nodes of out-edges from i; let β_i be the *predecessors* of i, which are the initial nodes of in-edges to i. We have $a_{i+} = |\alpha_i|$ and $a_{+i} = |\beta_i|$ for simple graphs, and $\alpha_i \equiv \beta_i$ for undirected graphs.

Where a distinction is relevant, (ij) denotes an element of $U \times U$, whereas (i, j) denotes a pair of nodes in U. Note that (ij) and (ji) are two distinct elements of $U \times U$ for digraphs, whereas they refer to the same element for undirected graphs. One can write $A_{ij} = \{(ij)\}$ if $a_{ij} = 1$, or $A_{ij} = \{(ij)_1, ..., (ij)_{a_{ij}}\}$ if $a_{ij} > 1$.

Finally, an edge incident to the same node i at both ends is called a *loop*, which can sometimes be a useful representation. Whether or not loops are included in the terms and notations defined above will be a matter of convention.

1.2.2 Motif

Let $M \subset U$. Let $G(M)$ be the subgraph *induced* by M, with node set M and edge set $A \cap (M \times M) = \{A_{ij} : i, j \in M\}$. For example, in the graph in Figure 1.1, the subgraph induced by $M = \{i_1, i_2, i_3, i_4\}$ has edges $\{(i_1i_2), (i_2i_3), (i_3i_1), (i_3i_4)\}$.

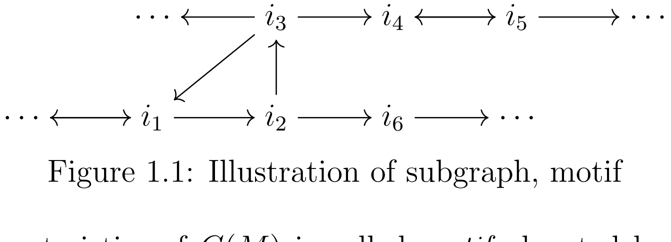

Figure 1.1: Illustration of subgraph, motif

The specific characteristics of $G(M)$ is called *motif*, denoted by $[M]$. The *order* of motif $[M]$ is $|M|$. For instance, $[i : a_{i+} = 3]$ defines the motif of a node with out-degree 3, when the node is considered as a subgraph $G(\{i\})$, such as i_3 in Figure 1.1. Whereas $[i, j : a_{ij}a_{ji} = 1]$ defines the motif of a node pair with 'mutual simple relationship', such as $G(\{i_4, i_5\})$ in Figure 1.1.

We have an *induced motif* $[M]$ if it can be determined based on the induced subgraph $G(M) = \big(M, A \cap (M \times M)\big)$ alone. Thus, $[i, j : a_{ij}a_{ji} = 1]$ is an induced motif, but not $[i : a_{i+} = 3]$ which requires the knowledge of $A \cap (\{i\} \times U)$ in addition. Some examples of low-order induced motifs in undirected graphs are given in Figure 1.2.

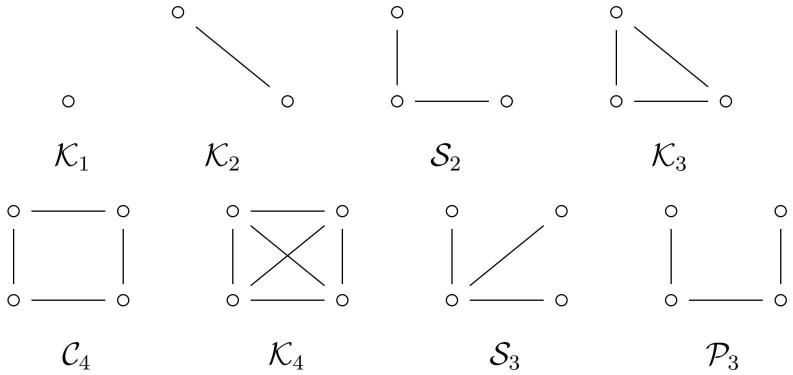

Figure 1.2: Node ($\mathcal{K}_1$), 2-clique ($\mathcal{K}_2$), 2-star ($\mathcal{S}_2$), 3-clique (triangle, $\mathcal{K}_3$), 4-cycle ($\mathcal{C}_4$), 4-clique ($\mathcal{K}_4$), 3-star ($\mathcal{S}_3$) and 3-path ($\mathcal{P}_3$), as induced motifs in undirected graphs.

A *component* of an undirected graph is an induced subgraph in which any two nodes are connected to each other by paths, and which is connected to no other nodes in the rest of the graph. It is an example of motif that is not induced. Let $|M| = K$, for $M \subset U$. Since whether or not $G(M)$ is a component in the graph depends on $A \cap \big(M \times U \cup U \times M\big)$, not just $A \cap (M \times M)$, it is not an induced motif.

As another example, a shortest path over ordered M is not an induced motif generally. Take ordered $M = \{i_1, i_2, i_3, i_4\}$ in Figure 1.1, whether or not $\{(i_1 i_2), (i_2 i_3), (i_3 i_4)\}$ is a shortest path (from i_1 to i_4 in G) depends also on $A \setminus (M \times M)$. Whereas if $M = \{i_1, i_2, i_3\}$, then $\{(i_1 i_2), (i_2 i_3)\}$ is a shortest path iff $(i_1 i_3) \notin A$, based on $G(M)$ alone.

1.2.3 Graph parameter

Let $y(M)$ be a function of $G(M)$. Let Ω contain all the relevant node sets M. The *graph total* of $y(M)$ over Ω is given by

$$\theta = \sum_{M \in \Omega} y(M) \tag{1.1}$$

It is said to be a *q-th order* graph total, if $|M| \equiv q$ for any $M \in \Omega$. The definition (1.1) applies to valued graphs, if $y(M)$ incorporates the values associated with $G(M)$.

We refer to arbitrary functions of $\{y(M) : M \in \Omega\}$ as *graph parameters.* A graph parameter is said to be of the *q-th order*, if $|M| \equiv q$ for any $M \in \Omega$. A graph parameter defined over Ω can involve different motifs. For example, let $|M| \equiv 3$ for all triplets from U, let θ_3 be the number of triangles in G, and θ_2 the number of 2-stars, both as illustrated in Figure 1.2. Let $\mu = \theta_3/(\theta_2 + \theta_3)$, which is a graph parameter of order 3.

The graph total (1.1) counts the motif of interest, denoted by $\mathfrak{m}$, when

$$y(M) = \mathbb{I}([M] = \mathfrak{m})$$

It may be convenient to let Ω denote the set of motifs of interest directly, such that each $\kappa \in \Omega$ is an *occurrence* of $\mathfrak{m}$ in G. For example, if $\mathfrak{m}$ is triangle in digraphs, then $\kappa \in \Omega$ is a distinct triangle in the digraph G, like the one corresponding to $G(M)$ with $M = \{i_1, i_2, i_3\}$ in Figure 1.1. A graph total over Ω can then be given as

$$\theta = \sum_{\kappa \in \Omega} y_\kappa$$

1.2.4 Examples of graph parameters

Each $M = \{i\}$ corresponds to a node $i \in U$. Each $M = \{i, j\}$ of a distinct pair of nodes (i, j) is called a *dyad.* Each $M = \{i, j, h\}$ of three distinct nodes is called a *triad.*

N = order (of G) Let $y(i) \equiv 1$, for $i \in U$. Then, $\theta = |U| = N$.

N_d = number of degree-d nodes Let $y(i) = \mathbb{I}(d_i = d)$, for $i \in U$.

R = size (of G) Let $y(i, j) = a_{ij}$ for $i, j \in U$ in directed graphs, or $a_{ij}/2$ in undirected graphs. Then, $\theta = R$. Meanwhile, for any undirected graph, we have

$$2R = \sum_{i \in U} d_i = \sum_{d=1}^{D} dN_d \qquad \text{where} \quad D = \max_{i \in U} d_i$$

and the 2nd-order total R is given as a graph parameter in terms of 1st-order totals N_d.

Adjacent pairs Let $y(i, j) = \mathbb{I}(a_{ij} + a_{ji} > 0)$ indicate whether i and j are adjacent. The corresponding ratio $\theta/\binom{N}{2}$ is a graph parameter as a measure of *immediacy.* A graph of minimum immediacy consists of only isolated nodes; a graph of maximum immediacy is a *clique*, where every pair of distinct nodes are adjacent.

Mutual relationships Let $y(i, j) = \mathbb{I}(a_{ij}a_{ji} > 0)$ indicate whether i and j have reciprocal edges between them, in which case their relationship is *mutual.*

Undirected triads For undirected simple graphs, Frank (1981) shows that there exists an explicit relationship between the mean and variance of the degree distribution and the triads of the graph. The numbers of triads of size 3, 2 and 1 are, respectively,

$$\begin{aligned}
\theta_{3,3} &= \sum_{i<j<h \in U} a_{ij}a_{jh}a_{ih} \\
\theta_{3,2} &= \sum_{i<j<h \in U} a_{ij}a_{ih}(1 - a_{jh}) + a_{ij}a_{jh}(1 - a_{ih}) + a_{ih}a_{jh}(1 - a_{ij}) \\
\theta_{3,1} &= \sum_{i<j<h \in U} a_{ij}(1 - a_{jh})(1 - a_{ih}) + a_{ih}(1 - a_{ij})(1 - a_{jh}) + a_{jh}(1 - a_{ij})(1 - a_{ih})
\end{aligned}$$

Let $\mu = \sum_{d=1}^{N} dN_d/N = 2R/N$ and $\sigma^2 = Q/N - \mu^2$, where $Q = \sum_{d=1}^{N} d^2 N_d$. We have

$$R = \frac{1}{N-2}\big(\theta_{3,1} + 2\theta_{3,2} + 3\theta_{3,3}\big) \qquad Q = \frac{2}{N-1}\big(\theta_{3,1} + N\theta_{3,2} + 3(N-1)\theta_{3,3}\big)$$

Transitivity For an undirected simple graph, $\theta_{3,3}$ above is the total number of triangles. Triangles are related to equivalence relationship: for a relationship (represented by an edge) to be transitive, any three connected nodes must form a triangle. It follows that transitivity is the case iff $\theta_{3,2} = 0$ over all connected triads. When $\theta_{3,2} \neq 0$, one can measure the extent of transitivity by the graph parameter $\theta_{3,3}/(\theta_{3,3} + \theta_{3,2})$.

Directed triangles In digraphs, let $z(j,i,h) = a_{ji}a_{ih}a_{hj}$ for ordered (j,i,h) be the count of *strongly connected* triangles from j via i and h back to j. Let $\tilde{M}$ contain all the possible orderings of a triad M, (i,j,h), (i,h,j), (j,i,h), (j,h,i), (h,i,j) and (h,j,i). The number of strongly connected triangles in a digraph is given by (1.1), where

$$y(M) = \sum_{(i,j,h)\in\tilde{M}} z(i,j,h)/3$$

Order-K component Let $|M| = K$ for $M \subseteq U$. Let $y(M) = 1$ if $[M]$ is a component and $y(M) = 0$ otherwise. The corresponding θ_K by (1.1) is the number of components of order K. The number of components (of unspecified order) is the graph parameter given by $\theta = \sum_{K=1}^{N} \theta_K$. At one end, where $A = \emptyset$, there are no edges at all in the graph, we have $\theta = N = \theta_1$ and $\theta_K = 0$ for $K > 1$. At the other end, where there exists a path between any two nodes, we have $\theta = \theta_N = 1$ and $\theta_K = 0$ for $K < N$.

Trees in a forest In an undirected simple graph, $[M]$ is a *tree* if the number of edges in $G(M)$ is $|M| - 1$. The graph is a *forest*, if every component of it is a tree. The total number of trees in a forest-graph is the graph parameter $\theta = N - R$.

Cliques The subgraph induced by a clique (of nodes) is said to be *complete*, where there exists an edge between any two nodes. A clustered population U can be represented by a graph, where each cluster is a clique, and two nodes i and j are adjacent iff they belong to the same cluster. Let edge stand for specified kinship, and θ the graph total of cliques. As a demographic mobility measure, θ/N varies between $1/N$ and 1, whereas the immediacy parameter $R/\binom{N}{2}$ varies between 1 and 0 in the other way.

Geodesic Let an undirected simple graph G be connected. An ordered set M is a geodesic if it consists of the nodes along a shortest path. Let Ω contain all such node sets, where $y(M) = |M| - 1$ is the geodesic length for $M \in \Omega$. As a closeness centrality measure, let θ be the harmonic mean of $y(M)$ over Ω, whose minimum value is 1 if G is complete.

1.3 Observation procedure

Compared to how samples are selected from finite populations, making use of the edges is a defining feature of selecting subgraphs from real graphs. The way in which the edges are used to drive subgraph selection is called the *observation procedure (OP)*.

Given en *initial sample (of nodes)* $s_0 \subset U$, the incident edges that are immediately available for any OP are either in $\alpha(s_0) = \bigcup_{i \in s_0} \alpha_i$ or $\beta(s_0) = \bigcup_{i \in s_0} \beta_i$, where $\alpha(s_0) = \beta(s_0)$ for undirected graphs.

The sample s_0 can as well be given as the nodes that are incident to a subset of edges from A. Initial sampling of edges may be useful if the graph is known but too large to be counted, or if it is more readily accessible via the edges.

Below we describe first the basic OPs given $s_0 \subset U$, and then the more complex ones that may be iterative, adaptive or probabilistic.

1.3.1 Basic OPs

An OP is *induced* when A_{ij} is observed iff both $i \in s_0$ and $j \in s_0$, or *incident-reciprocal* when A_{ij} and A_{ji} are both observed if either $i \in s_0$ or $j \in s_0$. Moreover, for digraphs, an incident non-reciprocal procedure is *forward* when A_{ij} is observed if $i \in s_0$, or *backward* when A_{ij} is observed if $j \in s_0$. It is convenient to specify the observed edges as

$$A_s = A \cap s_{\text{ref}}$$

via a *reference set*, denoted by s_{ref}, which explicates the parts of the *adjacency matrix* $[a_{ij}]$ that are observed given s_0 and the OP.

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

$$G: \qquad a \qquad b \longrightarrow c \longrightarrow d$$

Figure 1.3: Basic OP applied to $s_0 = \{b, d\}$ in G.

Take the graph G in Figure 1.3 as an example. Let the rows and columns of $[a_{ij}]$ be in the order a, b, c, d. Given $s_0 = \{b, d\}$, the reference set s_{ref} is marked by $\boxed{\cdot}$ for each basic OP. We observe none of the edges in G if the OP is induced, the edge (bc) if it is incident forward, or (cd) if incident backward, or both (bc) and (cd) if incident reciprocal.

Let π_i and π_{ij} be the inclusion probabilities *in* s_0. Let $\bar{\pi}_i = \Pr(i \notin s_0)$ be the *exclusion probability* of i, and $\bar{\pi}_{ij} = \Pr(i \notin s_0, j \notin s_0)$ that of $i, j \in U$. Similarly, $\bar{\pi}_M$ for any node set M. Let $\pi_{(ij)}$ be the probability of $(ij) \in A_s$ after applying the OP to s_0, and $\pi_{(ij)(kl)}$ that of the two edges (ij) and (kl) jointly.

(i) Induced, $s_{\text{ref}} = s_0 \times s_0$. Both $(ij) \in s_{\text{ref}}$ and $(ji) \in s_{\text{ref}}$ iff $i \in s_0$ and $j \in s_0$. Thus, $\pi_{(ij)} = \pi_{ij}$ and $\pi_{(ij)(kl)} = \pi_{ijkl}$.

(ii) Incident-forward, $s_{\text{ref}} = s_0 \times U$. We have $(ij) \in s_{\text{ref}}$ iff $i \in s_0$. Thus, $\pi_{(ij)} = \pi_i$ and $\pi_{(ij)(kl)} = \pi_{ik}$.

(iii) Incident-reciprocal, $s_{\text{ref}} = s_0 \times U \cup U \times s_0$. We have $(ij) \notin s_{\text{ref}}$ iff $i \notin s_0$ and $j \notin s_0$. Thus, $\pi_{(ij)} = 1 - \bar{\pi}_{ij}$ and $\pi_{(ij)(kl)} = 1 - \bar{\pi}_{ij} - \bar{\pi}_{kl} + \bar{\pi}_{ijkl}$.

From now on, let incident mean incident-forward in case of directed graphs, unless otherwise specified. For undirected graphs, one can always refer to incident-forward, incident-backward or incident-reciprocal all simply as incident.

1.3.2 Multiplicity

There exists generally a *multiplicity* in terms of access to any given motif in a graph. The multiplicity can be modified by the OP, in addition to the design of s_0.

Consider an epidemiology study in the population U, where a *case* is a person who would be positive when subjected to a diagnostic test. For any $i \in U$, let $y_i = 0$ if it is not a case, $y_i = 1$ if it is a hospitalised case, and $y_i = 2$ if it is a non-hospitalised case. Let $N = |U|$. Let $\mu = \sum_{i \in U} \mathbb{I}(y_i > 0)/N$ be the 1st-order graph parameter of interest.

Study-I Let s_0 be an initial sample of cases, who are receiving treatment at any of a sample of hospitals. Next, all the individuals that have been in contact with anyone in s_0 during the last fortnight are tracked down and tested.

Study-II Let s_0 be a simple random sample from U. Let s_0' be the subsample of cases, where $s_0' \subseteq s_0$. All the individuals that have been in contact with anyone in s_0' during the last fortnight are tracked down and tested.

Study-III Select and test one person $i_0 \in U$. Next, select-track-test randomly one among all the individuals that have been in contact with i_0 during the last fortnight, denoted by i_1. Repeat for i_{t+1} given i_t, where $t \geq 1$, till a fixed number n is reached.

Let $G = (U, A)$ be undirected, where $a_{ij} = 1$ if two persons $i, j \in U$ have had contact during the last fortnight, and $a_{ij} = 0$ otherwise. For each $i \in U$, let (d_{0i}, d_{1i}, d_{2i}) be the numbers of its adjacent nodes with $y = 0, 1, 2$, respectively, where $d_i = d_{0i} + d_{1i} + d_{2i}$.

In Study-I, s_0 is a single-stage cluster sample from $U_1 = \{i \in U : y_i = 1\}$, and the one-step tracing from each node in s_0 amounts to the incident OP. In particular, a node $j \in U_2 = \{i \in U : y_i = 2\}$ is selected, if $d_{1j} > 1$ and $\beta_j \cap s_0 \neq \emptyset$. Whenever $|\beta_j \cap s_0| < d_{1j}$, there are nodes outside of s_0, which could lead to j if the selection procedure is repeated. One may refer to d_{1j} as the multiplicity of $j \in U_2$ under the set-up of Study-I.

In Study-II, the incident OP is said to be *adaptive*, because tracing from $i \in s_0$ is implemented only if $y_i > 0$, such that it depends on the values associated with the graph in addition. The multiplicity of any $i \in U_1 \cup U_2$ is $1 + d_{1i} + d_{2i}$ in this set-up.

In Study-III, we have $s_0 = \{i_0\}$, where i_0 may or may not have a known probability of selection. The incident OP involves now *subsampling* of a single edge among all those incident to i_0. The successive *states* $\{i_0, i_1, ..., i_n\}$ form a Markov chain. Provided the graph is connected, the multiplicity of any $i \in U$ is d_i in this set-up.

1.3.3 Multiwave

We refer to *network* as a set of connected nodes, which satisfy some specific conditions. Any node in the graph is then called a *(network) edge node*, if it does not belong to the network but is adjacent to at least one node in the network; an edge node separates the network from the rest nodes. Figure 1.4 illustrates for an epidemiology study, where each network is a group of cases connected by contacts, and an edge node is a noncase that has contact with at least one case. If the (network) condition is void, then a network would simply be a component of the graph, which has no edge nodes.

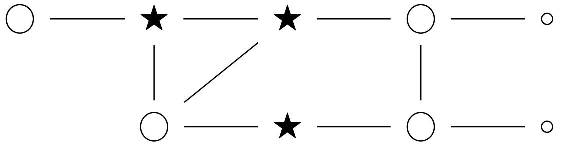

Figure 1.4: Illustration of network nodes (★), edge nodes (◯) and others (∘).

One can repeat the incident OP *wave by wave*. Such repetitions of the OP is said to be *network exhaustive*, if any network that intersects the initial sample s_0 will eventually be selected in its entirety. For Study-I or II, repeating the OP will eventually exhaust any case network that intersects s_0. Since s_0 is only selected from the hospitalised cases in Study-I, a network has no chance of being selected, if it does not contain any hospitalised cases; whereas every case network has a chance in Study-II. In Study-III, all the case networks will be exhausted, as $n \to \infty$, if the graph is connected.

A T-*wave* incident OP starting from s_0 is given as follows. For $t = 1, ..., T$, let

$$s_t = \alpha(s_{t-1}) \setminus \bigcup_{r=0}^{t-1} s_r$$

be the t-th wave sample of nodes. Let

$$s = \bigcup_{t=0}^{T-1} s_t$$

be the *seed sample*, to which the OP is applied. We have $s_{\text{ref}} = s \times U$. The OP is terminated, if $s_t = \emptyset$ for some $t < T$, in which case $s_{t+1} = \cdots = s_T = \emptyset$ as well. We have $T = 1$ in Study-I and II, and $T = n$ in Study-III.

1.3.4 Multigraph

In a multigraph, incident observation from i to j includes all the edges in A_{ij}. It can often be treated as incident OP in a valued simple graph, with a_{ij} from the original multigraph as the value attached to the edge (ij) in the simple graph, if $a_{ij} > 0$.

- In an undirected multigraph G, let each node correspond to a parent and each edge (ij) a child of i and j. Let A_{ii} represent the children of a single-parent i. Suppose one is interested in the structure of the parental networks.

- In a directed multigraph G, let each node correspond to a locality and each edge (ij) a telephone call from i to j. Let A_{ii} represent the calls within the locality. Suppose one is interested in the spatial connectivity based on telephone-call relationships.

- In a directed multigraph G, let each node correspond to a twitter account and each edge (ij) a retweet if i reposts Tweets from j. Suppose one is interested in the social dynamics or structure captured by retweets.

1.3.5 Multistage

Multistage sampling from finite populations is the case, when the ultimate sample of elements are selected in several steps, where the sampling units are different at every step. For example, sampling of hospitalised patients from an initial sample of hospitals corresponds to two-stage sampling.

A relational database is organised as a number of tables (or relations) of rows and columns. Each row (or record) in a table has a unique *key* for identification. Each table represents a distinct type of entity, whose columns contain the associated values (or attributes). The keys of one table can appear as attributes in other tables.

Let U be the union of all the collections of keys across the tables in a relational database. Let $(ij) \in A$ if the key j is an attribute in the table where i is an identification key, and i and j refer to different types of entities. The graph $G = (U, A)$ represents then the structure of the relational database, and selecting records from relational databases by keys can be represented as multistage OP in the graph G.

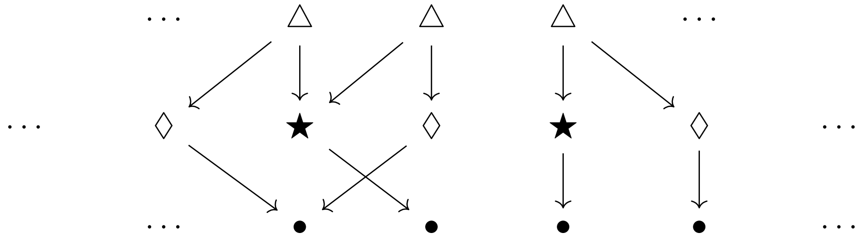

Figure 1.5: Relational database structure of entities $\triangle$, $\bigstar$, $\Diamond$ and $\bullet$.

Figure 1.5 illustrates a graph G representing the structure of a payment-account-holder relational database with 4 types of entity (the node set U): bank transaction ($\triangle$), payer bank account ($\bigstar$), receiver bank account ($\Diamond$), and bank account holder ($\bullet$). Let payer account and receiver account be two attributes of the table of transactions, such that edges exist from $\triangle$ to $\bigstar$ and to $\lozenge$ in the graph. Let holder be an attribute of the table of payer accounts and the table of receiver accounts, such that edges exist from $\bigstar$ or $\lozenge$ to $\bullet$. The edge set A can also be divided into four parts: from transactions to payer accounts ($\triangle\bigstar$), from transactions to receiver accounts ($\triangle\lozenge$), from payer accounts to holders ($\bigstar\bullet$), and from receiver accounts to holders ($\lozenge\bullet$).

Given a sample of transactions s_0, suppose one applies first incident OP to the edges from transactions to receiver accounts, then the edges from receiver accounts to holders. This can be viewed as a 3-stage OP in G, where holders are the ultimate sampling units, via the primary sampling units transactions and secondary sampling units receiver accounts. Subsampling of edges is also possible at the 2nd and 3rd stages.

1.3.6 Multilayer

In a *multipartite* graph, the node set U is partitioned into non-overlapping subsets, where edges only exist between nodes in different parts but not any nodes in the same part. In a *multilayer* graph, each edge in A is of a specific type (or dimension), where edges can exist between any pair of nodes in U. Multipartite or multilayer graphs can be given as valued graphs, where the partition of nodes can be defined via a value associated with each node, and that of edges via a value associated with each edge.

Formally, the graph of the relational database of payment-account-holder (Figure 1.5) can either be characterised as four-partite or four-dimensional. Suppose additional edges are introduced to represent an enriched structure of the payment-account-holder data, where $(ij), (ji) \in A$ if the two accounts i and j are affiliated with the same bank, and $(ij), (ji) \in A$ if the two holders i and j are mutual acquaintances. The graph is then multilayer and no longer simply multipartite.

Given a sample of transactions s_0, let the 1st-wave incident OP be applied to the edges from transactions to receiver accounts, and let the 2nd-wave incident OP be applied to the edges from receiver accounts to holders. In addition, let the incident OP be applied to the receiver accounts based on edges between them, and let the incident OP be applied to the holders based on edges between them. We shall refer to the incident OP as *multistage-multilayer*, since it is also applied to edges between the same type of entities.

The incident OP is simply *multilayer*, if it does not involve different types of nodes. For example, let U consist of individuals. Let A be two-dimensional, with edges between kins or colleagues. Applying separate incident OP to either type of edges yields a multilayer OP. In a way, multilayer OP resembles stratified sampling, in separate layers of edges rather than separate layers of the population, although multistage-multilayer OP in graphs graph is quite different to stratified sampling from finite populations.

1.4 Graph sampling strategy

Given a *population graph*, $G = (U, A)$, representing a population of units, U, and the connections (or links) between them, A, one may be interested in the structure of the connections, or the links may provide effectively access to the part of population that is 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 sampling.

Let the target of interest be given as a graph total θ or parameter μ. A *graph sampling strategy* consists of a graph sampling method and an accompanying estimator, just like a finite-population sampling strategy is consisted of a finite-population sampling method and an accompanying estimator. The properties of a graph sampling strategy is evaluated over hypothetical repeated sampling from the same population graph. Statistical inference with respect to the corresponding *known* sampling probabilities is said to be *design-based*. Design-based inference using a feasible graph sampling strategy is thus valid whatever the unknown properties of the population graph.

As much as graph sampling is universally applicable, it can be intricate to formulate a feasible strategy in various situations. Three elements are needed in any case.

I. Definition of sample graph, as a subgraph from the population graph.

II. Basis of inference, which can be different kinds of sampling probabilities associated with the given graph sampling method.

III. Eligible sample motifs. A motif observed in the sample graph is ineligible for estimation, if the required probabilities for inference cannot be calculated. Identification of eligible sample motifs is the key to any feasible graph sampling strategy.

1.4.1 Sample graph

Let $G = (U, A)$ be the population graph. A sampling method has generally two parts.

- Apply a given design to select an initial sample of nodes, $s_0 \subset U$.
- Apply a specified OP that makes use of the edges in A.

Denote by $G_s = (U_s, A_s)$ the *sample graph* with node sample U_s and edge sample A_s. The following definition of sample graph suffices in the sequel, where

$$G_s = (U_s, A_s) \qquad \text{with} \quad A_s = A \cap s_{\text{ref}} \quad \text{and} \quad U_s = s \cup \text{Inc}(A_s) \tag{1.2}$$

and s is the seed sample, to which the OP has been applied under graph sampling.

We do not consider separately sampling from graphs or value graphs. Generally, 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.

Simple random sampling (SRS) with or without replacement, Bernoulli sampling and Poisson sampling are some basic methods for sampling s_0 from U.

A sampling design is *symmetric* if $\pi_M = \Pr(M \subseteq s_0)$ for $M \subseteq U$ only depends on $|M|$ but is a constant of M otherwise, for all $1 \leq |M| \leq N$. SRS with or without replacement and Bernoulli sampling are all symmetric designs. SRS without replacement is the only symmetric design with fixed sample size of distinct elements.

The initial sample inclusion probabilities have simple expressions under Bernoulli sampling. Provided negligible sampling fraction of s_0, one may use Bernoulli sampling with probability $p = |s_0|/N$ to approximate any symmetric designs, or one may use Poisson sampling with probability π_i for initial sampling with unequal probability π_i, for $i \in U$. Monte Carlo simulation can be used to approximate the relevant initial sample inclusion probabilities under sampling without replacement.

The definition of sample graph by (1.2) includes the situation, where the initial sample s_0 is given as the nodes that are incident to a sample of edges directly selected from A. It is then possible for the OP to specify that no additional edges need to be sampled, in which case A_s contains the directly selected edges and $s = \emptyset$.

1.4.2 HT-estimator

A basic estimation method in graph sampling is the HT-estimator of a graph total (1.1). When the graph sample inclusion probability $\pi_{(M)}$ defined with respect to the reference set s_{ref} can be calculated for all $M \in \Omega$, the HT-estimator is given by

$$\hat{\theta} = \sum_{M \in \Omega} y(M) \delta_{[M]} / \pi_{(M)} \tag{1.3}$$

where $\delta_{[M]} = 1$ if the motif $[M]$ is observed in the sample graph G_s and 0 otherwise, and $\pi_{(M)} = \Pr(\delta_{[M]} = 1)$. The observation of $[M]$ requires not only $M \subseteq U_s$, but also it is possible to identify whether $[M]$ is the particular motif of interest. In particular, any induced motif $[M]$ is observed in G_s iff

$$M \times M \subseteq s_{\text{ref}} \tag{1.4}$$

Take 2-star in undirected simple graphs as an example. Suppose $s_{\text{ref}} = s_0 \times U$ by incident OP. To identify whether $[\{i, j, h\}]$ is a 2-star, s_{ref} needs to contain (ij), (ih) and (jh). Accordingly, $\pi_{(M)} = \Pr\big((ij) \in s_{\text{ref}}, (ih) \in s_{\text{ref}}, (jh) \in s_{\text{ref}}\big)$. An example where this is not the case is $i \in s_0$ and $j, h \in \alpha(s_0) \setminus s_0$, so that the observed part of the triad is a star, but one cannot be sure if $a_{jh} = 0$ in G, because $(jh) \notin s_{\text{ref}}$.

Given a graph parameter that is a function of one or several graph totals, a plug-in estimator is given by simply replacing each graph total in the definition of graph parameter by its HT-estimator, provided all these HT-estimators can be calculated.

1.4.3 Feasible graph sampling strategy

A feasible strategy may not exist for a given graph parameter, even though feasible strategies exist for many other graph parameters.

For example, suppose the sample inclusion probability of the nodes are only known up to a proportional constant, denoted by $\pi_i \propto c_i$ for $i \in U$, where c_i is known but π_i/c_i

is unknown because the graph order N is unknown. It is then only feasible to estimate the mean of the values associated with the nodes, but not their total.

Moreover, in a feasible strategy for a graph parameter, one may not be able to use *all* the relevant motifs that are observed in the sample graph.

For example, let incident OP be applied adaptively to a node only if the node is associated with a value above a specified threshold. All the nodes are then divided into two types of networks: (i) a network where all its member nodes have values above the threshold, (ii) a network consisting of a single node with a value not above the threshold. Let the adaptive OP be network exhaustive. The sample inclusion probability of a type-i network is the probability that it insects the initial sample s_0. An edge node to one or more type-i networks is a type-ii network. Its inclusion probability is unknown, unless one can be sure that the sample graph G_s contains all its adjacent type-i networks in G. It follows that one may not be able use an edge node even when it is in U_s.

For another example, given 3-wave incident OP in undirected graphs, any induced motif $[M]$ is observed if $M \times M \subseteq s_{\text{ref}}$. However, unless one has observed all the nodes that would have led to the observation of $[M]$, had any of them been selected in s_0, one would not be able to calculate $\pi_{(M)}$.

Forming a feasible strategy for the graph parameter of interest is generally not a trivial matter under graph sampling. It is a recurring theme in the following chapters.

Bibliographic notes

Neyman (1934) founded probability sampling from finite populations and the concept of sampling strategy. Horvitz and Thompson (1952) develop the basic theory of HT-estimation. Cochran (1977) describes the most common finite-population sampling techniques. A number of unconventional sampling methods are described in Thompson (2012), from either design or model-based perspective.

The adjacency matrix $\mathcal{A}$ is defined to be symmetric for undirected graphs. Denote by $\mathcal{D}$ the diagonal matrix of degrees. The Laplacian matrix $\mathcal{D} - \mathcal{A}$ sums to 0 by row and column, which is of central interest in Spectral Graph Theory (e.g. Chung, 1997).

Newman (2010) provides a comprehensive introduction to network analysis and many graph characteristics of interest. For a statistical approach to graph problems one may choose to model the entire valued graph as a random realisation; see e.g. Goldenberg et al. (2010) for a survey of statistical network models. Or, one may choose to exploit the variation over the possible sample graphs taken from a given real graph. Graph sampling theory deals with the valued graphs under the latter perspective.

For an early reference of graph sampling, Goodman (1961) studies the estimation of the number of mutual relationships in a special digraph, where $a_{i+} = 1$ for all $i \in U$.

Ove Frank makes many contributions to the graph sampling theory. See e.g. Frank (1977c, 1979, 1980b, 1981, 2011) for his own summary. However, the numerous works of Frank scatter over several decades, and are not easily appreciable as a whole. For instance, Frank derives results for *different* samples of nodes (Frank, 1971; 1977c; 1994), dyads (Frank, 1971; 1977a; 1977b; 1979) or triads (Frank, 1971; 1979). In other words, Ove Frank studies sampling of node sets or motifs, where a sample of motifs from the population of motifs is conceived in analogy to a sample s from the population U. But he never proposes a formal definition of sample graph, as a subgraph from the population graph, in which one may observe many different kinds of motifs. Or, Frank studies various characteristics of a graph, such as order (Frank, 1971; 1977c; 1994), size (Frank, 1971; 1977a; 1977b; 1979), degree distribution (Frank, 1971; 1980a), connectedness (Frank, 1971; 1978). But he never provides a structure of possible graph parameters which allows one to classify and contrast the different interests of study.

Zhang and Patone (2017) synthesise and extend the existing graph sampling theory. A definition of sample graph is proposed, where sampling is completely driven by the incident edges, whether the OP is multiwave, adaptive or random. This provides formally an analogy between sample graph as a sub-graph and sample as a sub-population. A structure of graph totals and parameters is given in terms of the order of motif. These extensions broaden the scope of investigation by graph sampling. The definition of sample graph (1.2) slightly generalises the definition adopted by Zhang and Patone (2017), which allows the OP to include random jumps to non-adjacent nodes.

Zhang and Patone (2017) consider T-wave snowball sampling, which is a probabilistic version of breadth-first search algorithms in graphs. The basis of inference is the graph sample inclusion probabilities of motifs. Other sampling probabilities may be needed, as will be seen later for targeted random walk sampling including random jumps, which is a probabilistic version of depth-first search algorithms in graphs.

Zhang (2021) provides an introduction to graph sampling mainly by examples.

Chapter 2

BIG sampling and IWE

For emphasis and distinction, denote by $\mathcal{B} = (F, \Omega; H)$ a *bipartite incidence graph (BIG)*, which is a simple digraph, where (F, Ω) form a bipartition of the node set $U = F \cup \Omega$, and each edge in H points from one node in F to another in Ω. No edge exists between any two nodes in F, nor in Ω, hence the graph is bipartite.

2.1 BIG Sampling

We define *BIG sampling (BIGS)* to be applying the incident OP to $s_0 \subset F$. The edges H are the *incidence links* from F to Ω. Given the initial (or seed) sample $s_0 \subset F$, the sample graph under BIGS, or *sample BIG*, is given by

$$\mathcal{B}_s = \big(s_0, \Omega_s; H_s\big) \tag{2.1}$$

with bipartite sample node set $U_s = s_0 \cup \Omega_s$, where $\Omega_s = \alpha(s_0)$ is a sample of nodes from Ω, and $s_{\text{ref}} = s_0 \times \Omega$ such that $H_s = H \cap s_{\text{ref}} = H \cap (s_0 \times \Omega)$.

We use $h, i, j, ...$ to enumerate F, which are referred to as the *sampling units* and assumed to have known initial sample inclusion probabilities, such as π_i, π_{ij}. We use $\kappa, \ell, ...$ to enumerate Ω, which are referred to as the *study units*. It is assumed that the graph total (or parameter) that is of interest is defined over Ω, e.g. by (1.1). Generally, F is assumed to be known, while both H and Ω may be unknown.

The basis of inference under BIGS is the sample inclusion probabilities. Despite the known initial sample inclusion probabilities, the study-sample inclusion probabilities, such as $\pi_{(\kappa)} = \Pr(\kappa \in \Omega_s)$ and $\pi_{(\kappa\ell)} = \Pr(\kappa \in \Omega_s, \ell \in \Omega_s)$, can only be calculated provided the knowledge of their *ancestors* in F (or their multiplicity), which are

$$\{\beta_\kappa : \kappa \in \Omega_s\} \quad \text{and} \quad \beta(\Omega_s) = \bigcup_{\kappa \in \Omega_s} \beta_\kappa$$

and will be referred to as the *ancestry knowledge*. The study units Ω_s are eligible for estimation, provided the ancestry knowledge in addition to the sample graph $\mathcal{B}_s$.

As additional regularity conditions for design-based inference of the graph total or parameter, we require $\pi_i > 0$ for each $i \in F$, and $\Omega = \alpha(F) = \bigcup_{i \in F} \alpha_i$.

Consider BIGS from $\mathcal{B}$ in Figure 2.1, where $F = \{i_1, i_2, i_3, i_4\}$, $\Omega = \{\kappa_1, \kappa_2, \kappa_3\}$ and $H = \{(i_1\kappa_1), (i_2\kappa_1), (i_2\kappa_2), (i_3\kappa_3)\}$. Suppose $s_0 = \{i_1, i_3\}$. We have $\Omega_s = \{\kappa_1, \kappa_3\}$ and $H_s = \{(i_1\kappa_1), (i_3\kappa_3)\}$, and the sample graph $\mathcal{B}_s = (s_0, \Omega_s; H_s)$ by (2.1). The ancestry knowledge consists of $\beta_{\kappa_1} = \{i_1, i_2\}$, $\beta_{\kappa_3} = \{i_3\}$ and $\beta(\Omega_s) \setminus s_0 = \{i_2\}$.

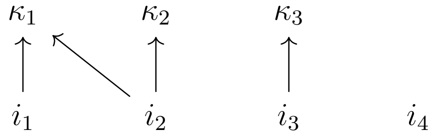

Figure 2.1: BIGS from $\mathcal{B}$

Finite-population sampling (FPS) is a special case of BIGS. Element sampling as BIGS is illustrated to the left in Figure 2.2, where $F = \Omega = U$, where $(i\kappa) \in H$ iff i and κ refer to the same population element. Cluster sampling as BIGS is illustrated to the right, where F consist of the m_F population clusters and $\Omega = U$ the elements that are nested in the clusters, where $(i\kappa) \in H$ iff element κ belongs to cluster i.

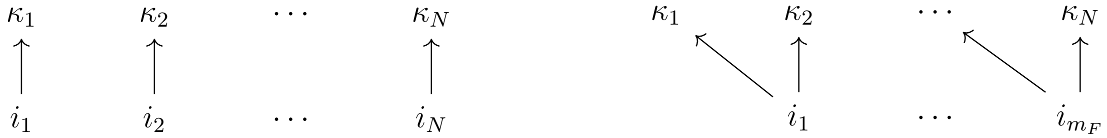

Figure 2.2: Population element sampling (left) and cluster sampling (right)

The knowledge of ancestry is guaranteed under FPS, because the mapping from F to Ω is one-one or one-many. The mapping can be many-one or many-many generally under BIGS, such that the ancestry knowledge needs to be required explicitly.

2.2 Incidence weighting estimator

Let $\theta = \sum_{\kappa \in \Omega} y_\kappa$ be the total of interest, where y_κ is a constant associated with node κ, for $\kappa \in \Omega$. Given the sample graph $\mathcal{B}_s$, let $\{W_{i\kappa}; (i\kappa) \in H_s\}$ be the *incidence weights* of the sample edges, where $W_{i\kappa} \equiv 0$ if $(i\kappa) \notin H_s$. The *incidence weighting estimator (IWE)* is given by

$$\hat{\theta} = \sum_{(i\kappa) \in H_s} W_{i\kappa} \frac{y_\kappa}{\pi_i} \tag{2.2}$$

Notice that, under BIGS, we have

$$\pi_i = \Pr(i \in s_0) = E(\delta_i) = \Pr\big((i\kappa) \in H_s\big) = \pi_{(i\kappa)}$$

where $\delta_i = 1$ or 0 indicates if $i \in s_0$ or not. The definition (2.2) allows $W_{i\kappa}$ to vary with the sample BIG. The condition for unbiased IWE under repeated sampling and its associated variance are given below. Replacing π_i by $\pi_{(i\kappa)}$ in (2.2) would allow the OP to involve subsampling among A_{i+}, for each $i \in s_0$, for which the results below can easily be rephrased accordingly.

Theorem 2.1. *The IWE by (2.2) is unbiased for θ provided, for each $\kappa \in \Omega$,*

$$\sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = 1 \tag{2.3}$$

Proof. The expectation of $\hat{\theta}$ with respect to the sampling distribution of s_0 is given by

$$E(\hat{\theta}) = \sum_{i \in F} \frac{E(\delta_i)}{\pi_i} E(\sum_{\kappa \in \alpha_i} W_{i\kappa} y_\kappa | \delta_i = 1) = \sum_{\kappa \in \Omega} y_\kappa \sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = \theta$$

since $\pi_{(i\kappa)} = \pi_i$ under BIGS, and $\sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = 1$ by stipulation. □

Corollary 2.1. *When the weights in (2.2) are constant of sampling, denoted by $\omega_{i\kappa}$ for distinction, the IWE is unbiased for θ provided, for each $\kappa \in \Omega$,*

$$\sum_{i \in \beta_\kappa} \omega_{i\kappa} = 1 \tag{2.4}$$

Proposition 2.1. *The BIGS variance of an unbiased IWE is given by*

$$V(\hat{\theta}) = \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} (\Delta_{\kappa\ell} - 1) y_\kappa y_\ell \tag{2.5}$$

where π_{ij} be the second-order initial sample inclusion probability of $i, j \in F$, and

$$\Delta_{\kappa\ell} = \sum_{i \in \beta_\kappa} \sum_{j \in \beta_\ell} \frac{\pi_{ij}}{\pi_i \pi_j} E\big(W_{i\kappa} W_{j\ell} | \delta_i \delta_j = 1\big)$$

Proof. Given unbiased $\hat{\theta}$, we have $V(\hat{\theta}) = E(\hat{\theta}^2) - \theta^2$, where

$$\begin{aligned} E\big(\hat{\theta}^2\big) &= \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} y_\kappa y_\ell \sum_{i \in \beta_\kappa} \sum_{j \in \beta_\ell} E\Big(\frac{\delta_i \delta_j}{\pi_i \pi_j} W_{i\kappa} W_{j\ell}\Big) \\ &= \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} y_\kappa y_\ell \sum_{i \in \beta_\kappa} \sum_{j \in \beta_\ell} \frac{\pi_{ij}}{\pi_i \pi_j} E(W_{i\kappa} W_{j\ell} | \delta_i \delta_j = 1) \end{aligned}$$

since $W_{i\kappa} W_{j\ell} = 0$ if $\delta_i \delta_j = 0$ under BIGS, for any $(i\kappa), (j\ell) \in H$. The result follows now from taking the difference of $E(\hat{\theta}^2)$ and $\theta^2 = (\sum_{\kappa \in \Omega} y_\kappa)^2$. □

2.2.1 HT-estimator (HTE)

Let $\pi_{(\kappa)} = \Pr(\kappa \in \Omega_s)$ and $\pi_{(\kappa\ell)} = \Pr(\kappa \in \Omega_s, \ell \in \Omega_s)$ for $\kappa, \ell \in \Omega$. Under BIGS, we have

$$\begin{aligned} \pi_{(\kappa)} &= 1 - \bar{\pi}_{\beta_\kappa} = 1 - \Pr\big(\beta_\kappa \cap s_0 = \emptyset\big) \\ \pi_{(\kappa\ell)} &= 1 - \bar{\pi}_{\beta_\kappa} - \bar{\pi}_{\beta_\ell} + \bar{\pi}_{\beta_\kappa \cup \beta_\ell} \end{aligned}$$

where $\bar{\pi}_{\beta_k}$ is the exclusion probability of β_k in s_0, which is the probability that none of the ancestors of κ in $\mathcal{B}$ is included in the initial sample s_0. The knowledge of the out-of-sample ancestors $\beta_\kappa \setminus s_0$ is required to compute $\bar{\pi}_{\beta_\kappa}$. Similarly for $\bar{\pi}_{\beta_\kappa \cup \beta_\ell}$.

The classic HT-estimator (HTE) is given by

$$\hat{\theta}_y = \sum_{\kappa\in\Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}}$$

and its sampling variance is given by

$$V(\hat{\theta}_y) = \sum_{\kappa\in\Omega}\sum_{\ell\in\Omega}\Big(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)}\pi_{(\ell)}} - 1\Big)y_\kappa y_\ell$$

An unbiased variance estimator can be given by

$$\widehat{V}(\hat{\theta}_y) = \sum_{\kappa\in\Omega_s}\sum_{\ell\in\Omega_s}\Big(\frac{1}{\pi_{(\kappa)}\pi_{(\ell)}} - \frac{1}{\pi_{(\kappa\ell)}}\Big)y_\kappa y_\ell$$

The HTE is a special case of the IWE, where the weights $W_{i\kappa}$ satisfy

$$\sum_{i\in s_0\cap\beta_\kappa} \frac{W_{i\kappa}}{\pi_i} = \frac{1}{\pi_{(\kappa)}} \tag{2.6}$$

which are not constant of sampling if $|\beta_\kappa| > 1$, depending on how s_0 intersects β_κ. For BIGS from Figure 2.1, we have

$$W_{i_2\kappa_2} = \frac{\pi_{i_2}}{\pi_{(\kappa_2)}} \qquad \text{and} \qquad W_{i_3\kappa_3} = \frac{\pi_{i_3}}{\pi_{(\kappa_3)}}$$

by (2.6), since both κ_2 and κ_3 have only one ancestor in the BIG. Moreover,

$$\begin{cases} W_{i_1\kappa_1} = \pi_{(\kappa_1)}^{-1}\pi_{i_1} & \text{if } s_0\cap\{i_1,i_2\} = \{i_1\} \\ W_{i_2\kappa_1} = \pi_{(\kappa_1)}^{-1}\pi_{i_2} & \text{if } s_0\cap\{i_1,i_2\} = \{i_2\} \\ (W_{i_1\kappa_1}, W_{i_2\kappa_1}) = \pi_{(\kappa_1)}^{-1}\big(a\pi_{i_1},\ (1-a)\pi_{i_2}\big) & \text{if } s_0\cap\{i_1,i_2\} = \{i_1,i_2\} \end{cases}$$

The value a does not matter, since the coefficient of y_{κ_1} in (2.2) is $\sum_{i\in s_0\cap\beta_\kappa} W_{i\kappa}/\pi_i$. A numerical illustration of $W_{i\kappa_1}$ can be found in Table 2.1 (Section 2.3).

To see that the weights given by (2.6) satisfy the condition (2.3) generally, let ϕ_{s_κ} be the probability that the *initial sample intersection* is $s_\kappa = s_0\cap\beta_\kappa$ for $\kappa\in\Omega$, where

$$\pi_{(\kappa)} = \sum_{s_\kappa}\phi_{s_\kappa}$$

over all possible s_κ. Given (2.6), for any $\kappa\in\Omega$, we have then

$$\sum_{i\in\beta_\kappa} E(W_{i\kappa}|\delta_i = 1) = \sum_{i\in\beta_\kappa}\sum_{s_\kappa\ni i}\frac{\phi_{s_\kappa}}{\pi_i}W_{i\kappa} = \sum_{s_\kappa}\phi_{s_\kappa}\sum_{i\in s_\kappa}\frac{W_{i\kappa}}{\pi_i} = \sum_{s_\kappa}\frac{\phi_{s_\kappa}}{\pi_{(\kappa)}} = 1$$

Similarly, for the joint probability that the sample intersections for κ and ℓ are s_κ and s_ℓ, it can be shown that $\Delta_{\kappa\ell}$ in (2.5) reduces to $\pi_{(\kappa\ell)}/\pi_{(\kappa)}\pi_{(\ell)}$ given (2.6) and (2.3).

2.2.2 HH-type estimator

The Hansen-Hurwitz (HH) type estimator uses weights $\omega_{i\kappa}$ that are constant of sampling, which is given by

$$\hat{\theta}_z = \sum_{i \in s_0} \frac{z_i}{\pi_i} \qquad \text{and} \qquad z_i = \sum_{\kappa \in \alpha_i} \omega_{i\kappa} y_\kappa \tag{2.7}$$

where z_i is a constructed constant for each $i \in F$. It is unbiased given (2.4), since

$$\sum_{i \in F} \sum_{\kappa \in \alpha_i} \omega_{i\kappa} y_\kappa = \sum_{\kappa \in \Omega} y_\kappa \sum_{i \in \beta_\kappa} \omega_{i\kappa} = \sum_{\kappa \in \Omega} y_\kappa$$

in which case its associated sampling variance is given by

$$V(\hat{\theta}_z) = \sum_{i \in F} \sum_{j \in F} (\frac{\pi_{ij}}{\pi_i \pi_j} - 1) z_i z_j$$

For BIGS from Figure 2.1, we have now $\omega_{i_1\kappa_1} + \omega_{i_2\kappa_1} = 1$ and $\omega_{i_2\kappa_2} = \omega_{i_3\kappa_3} = 1$ by (2.4).

Notice that one only needs z_i for the nodes in s_0 in order to apply $\hat{\theta}_z$, but not for the nodes outside s_0. The HH-type estimator (2.7) defines actually a family of estimators, depending on the choice of $\omega_{i\kappa}$. In particular, the equal weights

$$\omega_{i\kappa} = |\beta_\kappa|^{-1}$$

are referred to as the *multiplicity weights*, and the corresponding IWE is referred to as the *multiplicity estimator*, denoted by $\hat{\theta}_{z\beta}$.

2.2.3 Priority-rule estimator

Birnbaum and Sirken (1965) describe an estimator based on a *prioritised subset* of H_s. For each $\kappa \in \Omega_s$, they let $I_{i\kappa} = 1$ if $i = \min\big(s_0 \cap \beta_\kappa\big)$ and 0 otherwise, depending on if i happens to be enumerated first in F among all the in-sample ancestors of κ. The *priority-rule estimator* based on $\{(i\kappa) : I_{i\kappa} = 1, (i\kappa) \in H_s\}$ is given by

$$\hat{\theta}_p = \sum_{(i\kappa) \in H_s} \frac{I_{i\kappa} \omega_{i\kappa} y_\kappa}{p_{i\kappa} \pi_i} \tag{2.8}$$

where

$$p_{i\kappa} = \Pr\big(I_{i\kappa} = 1 | (i\kappa) \in H_s\big) = \Pr\big(I_{i\kappa} = 1 | \delta_i = 1\big)$$

is the conditional probability that $(i\kappa)$ is prioritised given $(i\kappa) \in H_s$, and $\omega_{i\kappa}$ is the multiplicity weight for $\kappa \in \Omega$. Clearly, other priority rules or choices of $\omega_{i\kappa}$ are possible.

One can easily recognise $\hat{\theta}_p$ as a special case of IWE with

$$W_{i\kappa} = I_{i\kappa} \omega_{i\kappa} / p_{i\kappa}$$

It can satisfy the unbiasedness condition (2.3), provided $p_{i\kappa} > 0$ given $(i\kappa) \in H_s$, in which case $E(W_{i\kappa}|\delta_i = 1) = \omega_{i\kappa}$. Its variance follows from (2.5), where

$$\Delta_{\kappa\ell} = \sum_{i\in\beta_\kappa}\sum_{j\in\beta_\ell} \frac{\pi_{ij} p_{i\kappa,j\ell}}{\pi_i \pi_j p_{i\kappa} p_{j\ell}} \omega_{i\kappa}\omega_{j\ell}$$

and $p_{i\kappa,j\ell} = \Pr(I_{i\kappa} I_{j\ell} = 1|\delta_i\delta_j = 1)$, such that

$$V(\hat{\theta}_p) = \sum_{(i\kappa)\in H}\sum_{(j\ell)\in H} \Big(\frac{\pi_{ij} p_{i\kappa,j\ell}}{\pi_i \pi_j p_{i\kappa} p_{j\ell}} - 1\Big) \omega_{i\kappa}\omega_{j\ell} y_\kappa y_\ell$$

because $\sum_{i\in\beta_\kappa} \omega_{i\kappa} = 1$ for any $\kappa \in \Omega$. An unbiased variance estimator can be given by

$$\hat{V}(\hat{\theta}_p) = \sum_{(i\kappa)\in H_s}\sum_{(j\ell)\in H_s} \Big(\frac{\pi_{ij} p_{i\kappa,j\ell}}{\pi_i \pi_j p_{i\kappa} p_{j\ell}} - 1\Big) \frac{\omega_{i\kappa}\omega_{j\ell}}{\pi_{ij}} y_\kappa y_\ell$$

The priority probabilities $p_{i\kappa}$ and $p_{i\kappa,j\ell}$ depend on both the priority rule and the initial sampling design of s_0. For the priority rule $\min(s_0 \cap \beta_\kappa)$, let $d_{i(\kappa)} = \sum_{j\in F: j<i} \mathbb{I}\big((j\kappa) \in H\big)$ be the number of nodes with higher priority than i for each $\kappa \in \Omega$ and $i \in \beta_\kappa$. Suppose initial SRS without replacement of s_0, where $m = |s_0|$. Let $m_F = |F|$. We have

$$p_{i\kappa} = \binom{m_F - 1 - d_{i(\kappa)}}{m-1} \Big/ \binom{m_F - 1}{m-1}$$

The joint priority probability of $(i\kappa)$ and $(j\ell)$ given $\delta_i\delta_j = 1$ is

$$p_{i\kappa,j\ell} = \begin{cases} p_{i\kappa} & \text{if } \kappa = \ell, i = j \\ 0 & \text{if } \kappa = \ell, i \neq j \\ \binom{m_F-1-d_{i(\kappa,\ell)}}{m-1} / \binom{m_F-1}{m-1} & \text{if } \kappa \neq \ell, i = j \\ \binom{m_F-2-d_{i(\kappa),j(\ell)}}{m-2} / \binom{m_F-2}{m-2} & \text{if } \kappa \neq \ell, i \neq j \text{ and } |\beta_\kappa^i \cap \{j\}| + |\beta_\ell^j \cap \{i\}| = 0 \\ 0 & \text{if } \kappa \neq \ell, i \neq j \text{ and } |\beta_\kappa^i \cap \{j\}| + |\beta_\ell^j \cap \{i\}| > 0 \end{cases}$$

where β_κ^i is the subset ancestors of κ with higher priority than i, and $d_{i(\kappa,\ell)} = |\beta_\kappa^i \cup \beta_\ell^i|$ is the number of nodes in $\beta_\kappa \cup \beta_\ell$ with higher priority than i, and $d_{i(\kappa),j(\ell)} = |\beta_\kappa^i \cup \beta_\ell^j|$.

Note that the priority rule is not part of sampling; the sample graph $\mathcal{B}_s$ includes all the edges incident to every node in s_0. Had one applied subsampling by randomly selecting one of the edges incident to each i in s_0, the sample graph would have contained one and only one edge from each $i \in s_0$. Instead, the priority rule selects only one sample edge incident to each study unit in Ω_s for the purpose of estimation.

There is a possibility that a node i can be sampled but never prioritised, in which case $\hat{\theta}_p$ would be biased. For an extreme example, suppose a study unit κ is adjacent to all the nodes in F, then the last node in F can never be prioritised (for κ) according to the priority rule $\min(s_0 \cap \beta_\kappa)$, as long as $|s_0| > 1$. Generally, $\hat{\theta}_p$ is biased under this priority rule, provided there exists at least one κ in Ω with $|\beta_\kappa| > 1$, where

$$\Pr(|s_\kappa| > 1 \mid \kappa \in \Omega_s) = 1$$

such that the ancestor $i = \max(\beta_\kappa)$ has no chance of being prioritised when it is in s_0. The probability above depends on the ordering of nodes in F, as well as the initial sample size. Given any ordering of the nodes in F, as the initial sample size increases, it is possible for $\hat{\theta}_p$ to behave more erratically and become biased eventually.

2.3 Rao-Blackwellisation

Given an unbiased estimator, its conditional expectation given the minimal sufficient statistic is also an unbiased estimator, which is referred to as *Rao-Blackwellisation (RB)*. The RB-estimator has a variance that is smaller than the initial estimator, if the two differ in at least some samples. The improvement can be substantive in certain situations. If the minimal sufficient statistic is complete in addition, then the RB-estimator is the unique minimum-variance unbiased estimator (UMVUE).

The minimal sufficient statistic under BIGS is $\{(\kappa, y_\kappa) : \kappa \in \Omega_s\}$, or simply Ω_s as long as one keeps in mind that the y-values are associated constants. However, it is not complete, because y_κ can be arbitrary constants over Ω. It follows that the RB method generally does not lead to UMVUE in graph or finite-population sampling.

Let $\hat{\theta}$ be an unbiased IWE. Applying the RB method to $\hat{\theta}$ yields

$$\hat{\theta}_{RB} = E(\hat{\theta}|\Omega_s)$$

as an improved estimator, if the conditional variance $V(\hat{\theta}|\Omega_s)$ is positive. Since the HTE $\hat{\theta}_y$ is fixed conditional on Ω_s, it will remain unchanged. It is in principle possible to use the RB method to improve the efficiency of a non-HT estimator.

For an illustration using Figure 2.1, consider BIGS given $|s_0| = 1$ first. There are 4 distinct initial samples leading to 4 distinct Ω_s, such that $V(\hat{\theta}|\Omega_s) = 0$ and $\hat{\theta}_{RB} = \hat{\theta}$ for any unbiased IWE. Next, given $|s_0| = 2$, there are 6 different initial samples, leading to 5 distinct Ω_s, where both $s_0 = \{i_1, i_2\}$ and $s_0' = \{i_2, i_4\}$ lead to the same $\Omega_s = \{\kappa_1, \kappa_2\}$, so that $\hat{\theta}_{RB} \neq \hat{\theta}$ given $\Omega_s = \{\kappa_1, \kappa_2\}$, if $\hat{\theta}(s_0) \neq \hat{\theta}(s_0')$. For the HH-type estimator $\hat{\theta}_z$ by (2.7), we have

$$\hat{\theta}_z(s_0) = \frac{\omega_{i_1\kappa_1}}{\pi_{i_1}} y_{\kappa_1} + \frac{\omega_{i_2\kappa_1}}{\pi_{i_2}} y_{\kappa_1} + \frac{\omega_{i_2\kappa_2}}{\pi_{i_2}} y_{\kappa_2} \neq \hat{\theta}_z(s_0') = \frac{\omega_{i_2\kappa_1}}{\pi_{i_2}} y_{\kappa_1} + \frac{\omega_{i_2\kappa_2}}{\pi_{i_2}} y_{\kappa_2}$$

$$\hat{\theta}_{zRB} = \frac{p(s_0)}{p(s_0) + p(s_0')} \cdot \frac{\omega_{i_1\kappa_1}}{\pi_{i_1}} y_{\kappa_1} + \frac{\omega_{i_2\kappa_1}}{\pi_{i_2}} y_{\kappa_1} + \frac{\omega_{i_2\kappa_2}}{\pi_{i_2}} y_{\kappa_2}$$

The calculation required of RB may be infeasible, if the conditional sample space of s_0 given Ω_s is large, or if the initial sampling distribution $p(s_0)$ is not fully specified. Nevertheless, the reasoning can offer valuable insights, as discussed below.

2.3.1 HT-type estimator

The HTE is based on sample-dependent weights $W_{i\kappa}$, which satisfy the constraint (2.6). More generally, let

$$\eta_{s_\kappa} = \pi_{(\kappa)} \sum_{i \in s_\kappa} \frac{W_{i\kappa}}{\pi_i}$$

To satisfy the condition (2.3), for any $\kappa \in \Omega$, the weights must be such that

$$\sum_{s_\kappa} \phi_{s_\kappa} \eta_{s_\kappa} = \pi_{(\kappa)}$$

The HTE is the special case of $\eta_{s_\kappa} \equiv 1$. It is possible to assign η_{s_κ} that differs from 1 for different sample intersects s_κ subject to this restriction. Any such non-HT estimator may be referred to as a *HT-type estimator*. However, applying the RB method to a HT-type estimator would recover the HTE, because

$$E\Big(\sum_{\kappa \in \Omega_s} \sum_{i \in s_\kappa} \frac{W_{i\kappa}}{\pi_i} y_\kappa | \Omega_s\Big) = \sum_{\kappa \in \Omega_s} y_\kappa E\Big(\frac{\eta_{s_\kappa}}{\pi_{(\kappa)}} | \kappa \in \Omega_s\Big) = \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}} \sum_{s_\kappa} \frac{\phi_{s_\kappa}}{\pi_{(\kappa)}} \eta_{s_\kappa} = \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}}$$

Table 2.1: Weight $W_{i\kappa_1}$ that varies with $s_{\kappa_1} = s_0 \cap \beta_{\kappa_1}$ under BIGS from Figure 2.1 given SRS of s_0 with $|s_0| = 2$, where $\pi_i \equiv 1/2$, $\pi_{(\kappa_1)} = 5/6$ and $\pi_i/\pi_{(\kappa_1)} \equiv 3/5$.

	HTE			HT-type, $(\eta_{s_{\kappa_1}}, \eta_{s_{\kappa_1}}) = (2/3, 1)$		
s_{κ_1}	$\{i_1, i_2\}$	$\{i_1\}$	$\{i_2\}$	$\{i_1, i_2\}$	$\{i_1\}$	$\{i_2\}$
$W_{i_1\kappa_1}$	$a(3/5)$	$3/5$	–	a	$2/5$	–
$W_{i_2\kappa_1}$	$(1-a)(3/5)$	–	$3/5$	$1-a$	–	$3/5$

Table 2.1 illustrates numerically the HTE-weights $W_{i\kappa_1}$ under BIGS from Figure 2.1 given SRS of s_0 with $|s_0| = 2$, where $\pi_i \equiv 1/2$ and $\pi_{(\kappa_1)} = 5/6$ such that $\pi_i/\pi_{(\kappa_1)} \equiv 3/5$. Given $\eta_{s_\kappa} \equiv 1$, we have $W_{i_1\kappa_1} = 3/5$ if $s_{\kappa_1} = \{i_1\}$, $W_{i_2\kappa_1} = 3/5$ if $s_{\kappa_1} = \{i_2\}$, and $W_{i_1\kappa_1} + W_{i_2\kappa_1} = 3/5$ if $s_{\kappa_1} = \{i_1, i_2\}$. Next, as an example of HT-type estimator, let $\eta_{s_{\kappa_1}} = 2/3$ and $W_{i_1\kappa_1} = 2/5$ if $s_{\kappa_1} = \{i_1\}$, and $\eta_{s_{\kappa_1}} = 1$ and $W_{i_2\kappa_1} = 3/5$ if $s_{\kappa_1} = \{i_2\}$. Since $\phi_{s_{\kappa_1}} = 1/3$ if $s_{\kappa_1} = \{i_1\}$ or $\{i_2\}$ and $\phi_{s_{\kappa_1}} = 1/6$ if $s_{\kappa_1} = \{i_1, i_2\}$, we must have

$$\frac{1}{6}\,\eta_{s_{\kappa_1}} = \frac{5}{6} - \Big(\frac{1}{3}\Big)\Big(\frac{2}{3}\Big) - \frac{1}{3} = \frac{5}{18} \quad \Rightarrow \quad W_{i_1\kappa_1} + W_{i_2\kappa_1} = \frac{3}{5}\,\eta_{s_{\kappa_1}} = 1$$

if $s_{\kappa_1} = \{i_1, i_2\}$. These $W_{i\kappa_1}$ are given in the right half of Table 2.1.

There is another way of looking at the HTE in terms of the RB method. One may consider each $s_\kappa = s_0 \cap \beta_\kappa$ as a sampling unit (instead of the ones in F), which can be arranged given any s_0 and $\alpha(s_0)$. Let

$$z_{s_\kappa} = W_{s_\kappa} y_\kappa$$

be the corresponding constructed value via the incident edge between s_κ and $\kappa \in \Omega$, where every s_κ leads to only one study unit κ in this construction. Consider an estimator based on $\{z_{s_\kappa} : \kappa \in \Omega_s\}$ given by

$$\hat{\theta} = \sum_{\kappa \in \Omega} \mathbb{I}(\kappa \in \Omega_s) \frac{z_{s_\kappa}}{\phi_{s_\kappa}}$$

where $\phi_{s_\kappa} = \Pr(s_\kappa = s_0 \cap \beta_\kappa)$ as before. It is unbiased, provided

$$E(\hat{\theta}) = \sum_{\kappa \in \Omega} y_\kappa \pi_{(\kappa)} E\big(\frac{W_{s_\kappa}}{\phi_{s_\kappa}} | \kappa \in \Omega_s\big) = \sum_{\kappa \in \Omega} y_\kappa \quad \Leftrightarrow \quad E\big(\frac{W_{s_\kappa}}{\phi_{s_\kappa}} | \kappa \in \Omega_s\big) = \frac{1}{\pi_{(\kappa)}}$$

for every $\kappa \in \Omega$. It follows that the corresponding $\hat{\theta}_{RB}$ is just the HTE.

2.3.2 HH-type estimator

Consider the special case of many-one mapping from F to Ω in $\mathcal{B}$, where $|\alpha_i| \equiv 1$. Compare $\hat{\theta}_y$ and $\hat{\theta}_z$ given $|s_0| = 1$. Let p_i and $p_{(\kappa)} = \sum_{i \in \beta_\kappa} p_i$ be the respective *selection* probabilities of $i \in F$ and $\kappa \in \Omega$. We have $p_{ij} = p_i$ if $i = j$ and 0 if $i \neq j$, and $p_{(\kappa\ell)} = p_{(\kappa)}$ if $\kappa = \ell$ and 0 if otherwise, now that $|\alpha_i| \equiv 1$. We have

$$V\big(\sum_{i \in s_0} \frac{z_i}{p_i}\big) - V\big(\sum_{\kappa \in \Omega_s} \frac{y_\kappa}{p_{(\kappa)}}\big) = \sum_{\kappa \in \Omega} \Big(\sum_{i \in \beta_\kappa} \frac{\omega_{i\kappa}^2}{p_i} - \frac{1}{p_{(\kappa)}}\Big) y_\kappa^2 = 0$$

only if $\omega_{i\kappa} \equiv p_i / p_{(\kappa)}$ for $i \in \beta_\kappa$, given which we have $\hat{\theta}_z = \hat{\theta}_{zRB}$. The variance of any other $\hat{\theta}_z$ would be larger, as long as $\omega_{i\kappa}/p_i$ is not a constant over β_κ, because

$$E\big(\frac{z_i}{p_i} | \kappa \in \Omega_s\big) = \sum_{i \in \beta_\kappa} \frac{p_i w_{i\kappa} y_\kappa}{p_{(\kappa)} p_i} = \frac{y_\kappa}{p_{(\kappa)}}$$

and

$$V\big(\frac{z_i}{p_i} | \kappa \in \Omega_s\big) = y_\kappa^2 V\big(\frac{\omega_{i\kappa}}{p_i} | \kappa \in \Omega_s\big) > 0$$

Since the same holds for sampling s_0 from F with replacement draw-by-draw, where $|s_0| > 1$, it suggests the choice $\omega_{i\kappa} \propto \pi_i$ under sampling of s_0 without replacement, provided $\pi_{ij} \approx \pi_i \pi_j$ and $\pi_{(\kappa\ell)} \approx \pi_{(\kappa)} \pi_{(\ell)}$. This can make z_i/π_i more similar to each other over F, which is advantageous with respect to the anticipated mean squared error of $\hat{\theta}_z$, evaluated under the sampling design *and* a population model of z_i, according to a result by Godambe and Joshi (1965, Theorem 6.2).

To make z_i/π_i more similar to each other over F without the restriction $|\alpha_i| = 1$, one may consider setting $\omega_{i\kappa} < \omega_{j\kappa}$ if $|\alpha_i| > |\alpha_j|$, despite $\pi_i = \pi_j$, for $i \neq j \in F$, because there are more study units contributing to z_i than z_j. Thus, it may be reasonable to consider the *probability and inverse-degree adjusted (PIDA)* weights for $\hat{\theta}_z$, which is given by

$$\omega_{i\kappa} \propto \pi_i / |\alpha_i|^\gamma \tag{2.9}$$

subjected to the condition (2.4), where $\gamma > 0$ is a tuning constant of choice.

Denote by $\hat{\theta}_{z\alpha\gamma}$ the corresponding PIDA-IWE. The multiplicity estimator $\hat{\theta}_{z\beta}$ becomes a special case of $\hat{\theta}_{z\alpha\gamma}$ if $\gamma = 0$ and π_i is constant over F.

To apply the weights (2.9) with $\gamma \neq 0$, one needs to know $|\alpha_i|$ for all $i \in \beta_\kappa$ and $\kappa \in \Omega_s$, in addition to the usual ancestry knowledge. For instance, let F be the parents and Ω the children, one would need to obtain the number of children for the out-of-sample parents in $\beta(\Omega_s) \setminus s_0$, in addition to the out-of-sample parents themselves.

2.4 Illustrations

Let $y_\kappa \equiv 1$ for $\kappa \in \Omega$, such that the graph total by (1.1) is $\theta = |\Omega|$. BIGS can be used to estimate θ in many situations.

- Let F be the population and Ω the cases in an epidemiology study, where $(i\kappa) \in H$ if person i is in-contact with case κ, including when i and κ refer to the same person.
- Let F be the Twitter accounts and Ω the followers. Let α_i be the followers of $i \in F$. The number of distinct followers is neither $\sum_{i \in F} a_{i+}$ nor $|F|$.
- Let F be the products available in an online market place and Ω the buying customers. Let α_i be those who have purchased the product $i \in F$.

Consider the following IWEs:

- the HTE $\hat{\theta}_y$;
- the HH-type estimator $\hat{\theta}_{z\alpha\gamma}$ by (2.9), where $\hat{\theta}_{z\alpha 0}$ is the multiplicity estimator;
- the priority-rule estimator $\hat{\theta}_p$ by (2.8) with multiplicity weights $\omega_{i\kappa}$ and priority rule $\min(s_0 \cap \beta_\kappa)$, where F is arranged in random, ascending or descending order of a_{i+}, yielding three estimators, denoted by $\hat{\theta}_{pR}$, $\hat{\theta}_{pA}$ and $\hat{\theta}_{pD}$, respectively.

2.4.1 A numerical example

Consider sampling from the BIG below, with SRS of s_0 and $m = |s_0| = 2$.

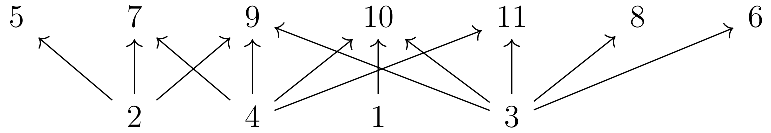

Figure 2.3: An example from Patone (2020)

The PIDA weights (2.9) for the HH-type estimator (2.7) is given in Table 2.2, as are the corresponding constructed z-values in (2.7). Under SRS of s_0 where $\pi_i \equiv |s_0|/m_F = 0.5$ with $m_F = |F|$, we have

$$V(\hat{\theta}_z) = m_F^2 (\frac{1}{m} - \frac{1}{m_F}) S_z^2 \quad \text{and} \quad S_z^2 = \sum_{i \in F} \frac{(z_i - \bar{z})^2}{m_F - 1} \quad \text{and} \quad \bar{z} = \sum_{i \in F} \frac{z_i}{m_F}$$

Table 2.2: PIDA weights and z-value given γ under BIGS from Figure 2.3.

γ	$\omega_{1,10}$	$\omega_{2,5}$	$\omega_{2,7}$	$\omega_{2,9}$	$\omega_{3,10}$	$\omega_{3,8}$	$\omega_{3,11}$	$\omega_{3,9}$	$\omega_{3,6}$	$\omega_{4,7}$	$\omega_{4,10}$	$\omega_{4,11}$	$\omega_{4,9}$
0	0.33	1	0.50	0.33	0.33	1	0.5	0.33	1	0.5	0.33	0.5	0.33
1	0.69	1	0.57	0.43	0.14	1	0.44	0.26	1	0.43	0.17	0.56	0.32
2	0.90	1	0.64	0.52	0.04	1	0.39	0.19	1	0.36	0.06	0.61	0.29
3	0.98	1	0.70	0.61	0.007	1	0.34	0.14	1	0.30	0.013	0.66	0.25

γ	z_1	z_2	z_3	z_4	S_z^2
0	0.33	1.83	3.17	1.67	1.34
1	0.69	2.00	2.83	1.48	0.81
2	0.91	2.16	2.61	1.32	0.60
3	0.98	2.31	2.48	1.23	0.57

where S_z^2 is the variance of z_i over F. It can be seen that the inverse-degree adjustment $1/|\alpha_i|^\gamma$ can heavily reduce the variance compared to the multiplicity weights.

The incident weights $W_{i\kappa}$ for the priority-rule estimator (2.8) vary with the initial sample s_0, as well as the ordering of F. We have $\tilde{F} = \{2, 4, 1, 3\}$ in Figure 2.3. Whereas we would have $\tilde{F} = \{3, 4, 2, 1\}$ if the nodes are arranged in the descending order of $|\alpha_i|$, or $\tilde{F} = \{1, 2, 4, 3\}$ in the ascending order of $|\alpha_i|$.

Table 2.3: Incident weights $W_{i\kappa}$ by priority rule given $\tilde{F} = \{2, 4, 1, 3\}$ in Figure 2.3.

s_0	$W_{1,10}$	$W_{2,5}$	$W_{2,7}$	$W_{2,9}$	$W_{3,10}$	$W_{3,8}$	$W_{3,11}$	$W_{3,9}$	$W_{3,6}$	$W_{4,7}$	$W_{4,10}$	$W_{4,11}$	$W_{4,9}$
$\{1,2\}$	0.33	1	0.5	0.33	-	-	-	-	-	-	-	-	-
$\{1,3\}$	0.33	-	-	-	0	1	0.5	0.5	1	-	-	-	-
$\{1,4\}$	0.33	-	-	-	-	-	-	-	-	0.75	0	0.75	1
$\{2,3\}$	-	1	0.5	0.33	0.5	1	0.5	0	1	-	-	-	-
$\{2,4\}$	-	1	0.5	0.33	-	-	-	-	-	0	1	0.75	0
$\{3,4\}$	-	-	-	-	0.5	1	0.5	0.5	1	0.75	0	0	0

The incidence weights given $\tilde{F} = \{2, 4, 1, 3\}$ are shown in Table 2.3. For instance, given $s_0 = \{1, 2\}$, we have $\alpha(s_0) = \{10, 5, 7, 9\}$, and all the sample edges in H_s are prioritised, since $|s_0 \cap \beta_\kappa| = 1$ for each $\kappa \in \alpha(s_0)$. Whereas given $s_0 = \{1, 3\}$, we have $\alpha(s_0) = \{10, 9, 11, 8, 6\}$, and all the sample edges are prioritised except that from 3 to 10, since $s_0 \cap \beta_{10} = \{1, 3\}$ and the edge from 1 to 10 is prioritised.

Table 2.4: Variance of IWE under BIGS from $\mathcal{B}$ above, $|s_0| = 2$.

	$\hat{\theta}_{z\alpha 0}$	$\hat{\theta}_{z\alpha 1}$	$\hat{\theta}_{z\alpha 2}$	$\hat{\theta}_{z\alpha 3}$	$\hat{\theta}_{pR}$	$\hat{\theta}_{pD}$	$\hat{\theta}_{pA}$	$\hat{\theta}_y$
Variance	5.37	3.25	2.41	2.28	3.06	2.55	6.32	3.98

Table 2.4 gives the sampling variances, which vary from 2.28 at the lowest to 6.32 at the highest. The choice of the estimator can matter quite much. The best HH-type estimator shown here has variance 2.28 and the best priority-rule estimator has variance 2.55, both of which are considerably more efficient than the HTE.

2.4.2 A simulation study

Two graphs $\mathcal{B} = (F, \Omega; H)$ and $\mathcal{B}' = (F, \Omega; H')$ are constructed for this illustration, which have the same F and Ω, where $|F| = 54$ and $|\Omega| = 310$. The two edge sets have the same number of edges, where $|H| = |H'| = 1200$, but different distributions of the out-degree $a_{i+} = |\alpha_i|$ over F, as shown in Figure 2.4. The distribution is relatively uniform over a small range of values in $\mathcal{B}$, but much more skewed and asymmetric in $\mathcal{B}'$.

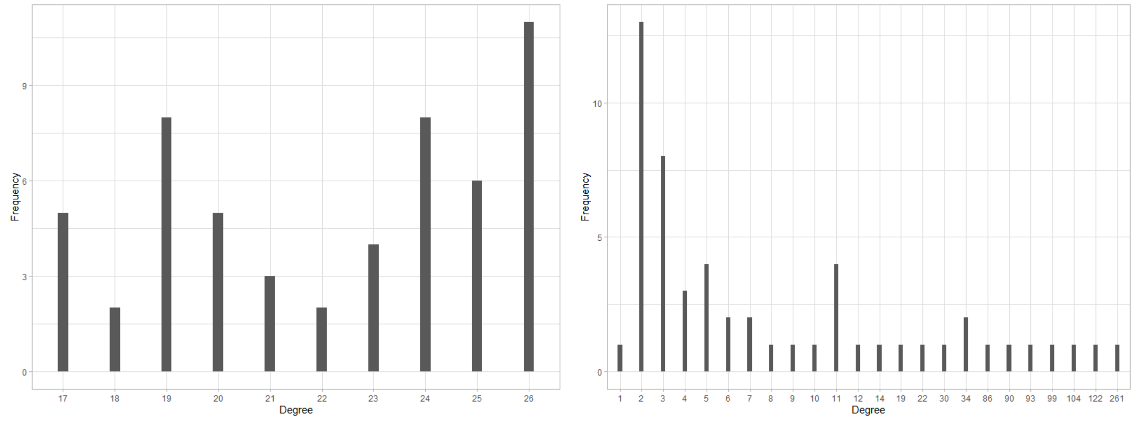

Figure 2.4: Distribution of out-degree a_{i+} in $\mathcal{B}$ (left) and $\mathcal{B}'$ (right).

Suppose SRS of s_0, where $m = |s_0| = 2, ..., 53$. Table 2.5 gives the relative efficiency (RE) of six other estimators against the HTE $\hat{\theta}_y$, for a selected set of initial sample sizes, each based on 10000 simulations of BIGS from either $\mathcal{B}$ or $\mathcal{B}'$. All the results are significant with respect to the simulation error.

Table 2.5: Relative efficiency of IWE for $\mathcal{B}$ and $\mathcal{B}'$, 10000 simulations.

	$\mathcal{B}$						$\mathcal{B}'$					
m	$\hat{\theta}_{z\alpha 0}$	$\hat{\theta}_{z\alpha 1}$	$\hat{\theta}_{z\alpha 2}$	$\hat{\theta}_{pR}$	$\hat{\theta}_{pA}$	$\hat{\theta}_{pD}$	$\hat{\theta}_{z\alpha 0}$	$\hat{\theta}_{z\alpha 1}$	$\hat{\theta}_{z\alpha 2}$	$\hat{\theta}_{pR}$	$\hat{\theta}_{pA}$	$\hat{\theta}_{pD}$
5	0.96	0.55	0.49	0.80	1.43	0.68	1.22	0.23	0.18	0.97	1.16	0.83
11	0.95	0.55	0.48	0.97	2.57	0.84	1.74	0.33	0.25	0.89	1.54	0.45
17	0.99	0.57	0.51	2.34	4.98	2.57	2.67	0.51	0.39	0.82	2.30	0.24
29	1.31	0.75	0.67	26.7	30.1	33.2	7.96	1.54	1.17	12.0	12.1	29.3

It can be seen in Table 2.5 that the priority-rule estimator is always dominated by some HH-type estimator in this simulation study. Moreover, from details omitted here, it is observed that all the three estimators $\hat{\theta}_{pR}$, $\hat{\theta}_{pA}$ and $\hat{\theta}_{pD}$ become biased given large enough initial sample size m, which happens at $m = 45$ for $\mathcal{B}$ where the maximum degree $|\beta_\kappa|$ is 10 over Ω, and $m = 46$ for $\mathcal{B}'$ where the maximum degree $|\beta_\kappa|$ is 9. Moreover, although the variance of any $\hat{\theta}_p$ initially decreases as m increases, the variance starts to increase with m once the latter is larger than a threshold value, somewhere between 10 and 30 in these simulations, so that the performance of $\hat{\theta}_p$ can deteriorate as the initial sample size increases long before it becomes biased.

The sampling variance of $\hat{\theta}_p$ is also affected by the ordering of the nodes in F. The variance tends to be lowest when F is arranged in descending ordering by $|\alpha_i|$, as long as the variance is decreasing with m, whereas ascending ordering tends to yield the largest variance. Without prioritisation, the value z_i is a constant of sampling given $\omega_{i\kappa}$. Due the randomness induced by the priority rule, z_i varies over different samples. A node with large $|\alpha_i|$ has a large range of possible z_i values. Placing such a node towards the end of the ordering tends to increase the sample variance of $\{z_i : i \in s_0\}$ due to prioritisation, compared to when the same node is placed towards the beginning of the ordering, because the rule $\min(s_0 \cap \beta_\kappa)$ favours the node in front of the other ancestors. This is a reason why descending ordering by $|\alpha_i|$ may work better than ascending ordering. However, one may not know $\{|\alpha_i| : i \in F\}$ in practice, in which case applying $\hat{\theta}_p$ given *ad hoc* ordering of F can be a haphazard business.

Given initial SRS, the different HH-type estimators based on the PIDA weights (2.9) differ only with respect to the use of $|\alpha_i|$ via the choice of γ. The equal-weights estimator $\hat{\theta}_{z\alpha 0}$ is the least efficient here, especially for $\mathcal{B}'$ where the distribution of $|\alpha_i|$ is more skewed. The differences between the other two estimators $\hat{\theta}_{z\alpha 1}$ and $\hat{\theta}_{z\alpha 2}$ are relatively small, compared to their differences to $\hat{\theta}_{z\alpha 0}$, so that a non-optimal choice of $\gamma \neq 0$ is less critical than simply setting $\gamma = 0$. Taken together, these results suggest that the extra effort that may be required to obtain $|\alpha_i|$ is worth considering in applications.

Finally, both $\hat{\theta}_{z\alpha 1}$ and $\hat{\theta}_{z\alpha 2}$ are more efficient than the HT-estimator $\hat{\theta}_y$ when m is small, whereas $\hat{\theta}_y$ improves more quickly as m becomes larger, especially for $\mathcal{B}'$. Generally, the difference between the two sampling variances can be given as

$$V(\hat{\theta}_z) - V(\hat{\theta}_y) = \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} \Big(\sum_{i \in \beta_\kappa} \sum_{j \in \beta_\ell} \frac{\pi_{ij}}{\pi_i \pi_j} \omega_{i\kappa} \omega_{j\ell} - \frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)} \pi_{(\ell)}} \Big) y_\kappa y_\ell$$

Thus, the RE between the two depends on the sampling fractions $|s_0|/|F|$ and $|\Omega_s|/|\Omega|$, as well as the respective inclusion probabilities of the nodes in F and Ω. The interplay between them is complex as it depends on the population edge set H. The matter is not really simplified in any helpful way under SRS of the initial sample here, where we have $\pi_i = m/m_F$ and $\pi_{ij} = m(m-1)/m_F(m_F - 1)$, for $m_F = |F|$ and any $i \neq j \in F$.

Bibliographic notes

Birnbaum and Sirken (1965) study the situation where patients are sampled indirectly via the hospitals from which they receive treatment. Insofar as a patient may be treated at more than one hospital, the patients are not nested in the hospitals like elements in clustered sampling. The knowledge of multiplicity (of the sampled patients) is necessary in order to calculate the HT-estimator. Indirect sampling as such can be naturally represented as sampling from BIG, where F consists of the hospitals and Ω the patients, with an edge between a hospital and each of its patients. Following the incident OP from the initial sample s_0 of hospitals, the knowledge of multiplicity requires the observation of β_κ for each patient $\kappa \in \Omega_s$ in addition.

Birnbaum and Sirken (1965) do not cast the problem in terms of graph sampling. They identify the condition (2.4) for unbiased HH-type estimator (2.7), although they only use the multiplicity weights $\omega_{i\kappa} = 1/|\beta_\kappa|$. Variations of the multiplicity estimator under other settings of indirect, network sampling are considered by Sirken (1970), Sirken and Levy (1974), Sirken (2004, 2005) and Lavalleè (2007). A modified multiplicity estimator is considered for adaptive cluster sampling (Thompson, 1990; 1991).

The priority-rule estimator is the third estimator considered by Birnbaum and Sirken (1965), who do not provide an expression of its sampling variance but indicate that it is unwieldy. The estimator seems to have vanished from the literature since then. Patone and Zhang (2020) provide the variance of $\hat{\theta}_p$ under SRS of s_0 and the priority rule $\min(s_0 \cap \beta_\kappa)$. Moreover, they show that the application of the priority-rule estimator may be a haphazard business, unless one is able to control the interplay between the ordering of F and the adopted priority-rule, and the estimator may become biased as the initial sample size increases and behave erratically long before that. It is unclear at this stage whether these shortcomings can be overcome in future.

Patone and Zhang (2020) formulate the unifying class of IWE for BIGS, and identify the general condition (2.3) for unbiased estimation, which allows for sample-dependent weight $W_{i\kappa}$, beyond the existing estimators that are based on sample-constant weights $\omega_{i\kappa}$ and the associated condition (2.4). In particular, the classic HTE is shown to be a special case of IWE. Other incidence weights given the condition (2.3) may emerge, beyond those examined in this chapter.

Rao-Blackwellisation (Rao, 1945; Blackwell, 1947) is unusual in the sampling theory, because the minimal sufficient statistic is not complete and the HTE is unchanged by it. A notable exception is the works related to adaptive sampling, whereby modifications of the HTE and multiplicity estimator are used; see e.g. Thompson (1990; 1991; 2006b), Dryver and Thompson (2005), Vincent and Thompson (2017).

The numerical example (Sec. 2.4.1) is worked out by Patone (2020), and the simulation study (Sec. 2.4.2) can be found in Patone and Zhang (2020).

Chapter 3

Strategy BIGS-IWE

The strategy BIGS-IWE can be applied to many sampling situations, whether or not these are originally described as a graph problem.

3.1 Applicability

For graph sampling that yields sample graph G_s from $G = (U, A)$ by (1.2), let $F \subseteq U$ be the initial sampling frame, consisting of the nodes in U which have positive probabilities of being included in the initial sample s_0. Let θ be the graph total given by (1.1), where Ω consists of induced motifs κ, each defined for a network of nodes $M(\kappa)$ or simply M as long as the context is clear that M and κ correspond to each other. In particular, the motifs κ can refer to single nodes, where $|M(\kappa)| \; \equiv 1$.

Let Ω_s be the motifs that are observed in the sample graph G_s, where $M \times M \subseteq s_{\text{ref}}$ for each $\kappa \in \Omega_s$. Let $\beta_\kappa \subseteq F$ be the subset of nodes, where

$$\Pr(\kappa \in \Omega_s | i \in s_0) = 1 \tag{3.1}$$

under graph sampling from G, for any $i \in \beta_\kappa$ and $\kappa \in \Omega$. That is, under graph sampling from G with associated F and Ω, the motif κ is observed in the sample Ω_s whenever a node i in β_κ is included in the initial sample s_0. Let $H = \bigcup_{\kappa \in \Omega} \beta_\kappa \times \kappa$.

Theorem 3.1. *The strategy BIGS-IWE defined for $\mathcal{B} = (F, \Omega; H)$ subjected to* (3.1) *and* (2.3) *is unbiased for θ defined by* (1.1)*, under graph sampling from $G = (U, A)$ defined by* (1.2)*, provided*

(i) $\forall \kappa \in \Omega$, we have $\beta_\kappa \neq \emptyset$ in $\mathcal{B}$, or equivalently $\bigcup_{i \in F} \alpha_i = \Omega$ in $\mathcal{B}$;

(ii) the OP of graph sampling from G ensures the ancestry knowledge of Ω_s in $\mathcal{B}$.

Proof. Given (i), every motif in Ω has a positive probability of being included in Ω_s with respect to BIGS from the constructed $\mathcal{B} = (F, \Omega; H)$. Given (ii), the IWE (2.2) can be defined with respect to BIGS from $\mathcal{B}$ by virtue of (3.1). Subjected to (2.3), the expectation of the IWE over repeated sampling from G is θ, by Theorem 2.1. □

Note that graph sampling based on the induced OP has difficulty satisfying (3.1) generally. For instance, let $\Omega = A$, such that a motif $\kappa = (ij)$ is observed in $\Omega_s = A_s$ iff $(i,j) \in s_0$, so that $\Pr(\kappa \in \Omega_s | i \in s_0) < 1$ for any $\kappa \in \Omega$. The edge set H would be empty in $\mathcal{B}$, which violates the condition (i) of Theorem 3.1.

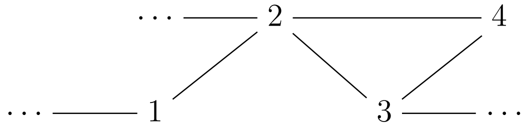

Figure 3.1: Illustration for Theorem 3.1.

To illustrate the condition (ii) of Theorem 3.1, suppose $F = U$ and let Ω contain all the 2-stars in an undirected graph G. Suppose the sample graph (1.2) is obtained by 2-wave incident OP given $s_0 \subset F$. For the 2-star motif κ of $M = \{1,2,3\}$ in Figure 3.1, we have $\beta_\kappa = \{1,2,3,4\}$ in $\mathcal{B}$ by construction (3.1). However, the condition (ii) is not satisfied by 2-wave incident OP in G, e.g. 4 cannot be identified to belong to β_κ if $s_0 \cap \beta_\kappa = \{1\}$ since the edge (34) is unobserved in the sample graph. How to apply the strategy BIGS-IWE in such situations will be discussed in Chapters 4 and 5.

3.2 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. Figure 3.2 presents the situation as graph sampling from $G = (U, A)$, where the node set U contains all the households ($\diamond$) and siblings ($\bullet$). Examples of sibling networks of order 1, 2 and 3 are given in Figure 3.2. The edges from $\diamond$ to $\bullet$ represent the access links from households to siblings, while the edges between two $\bullet$-nodes stand for the sibling relationship, such that G is a multiplayer (2-dimensional) graph.

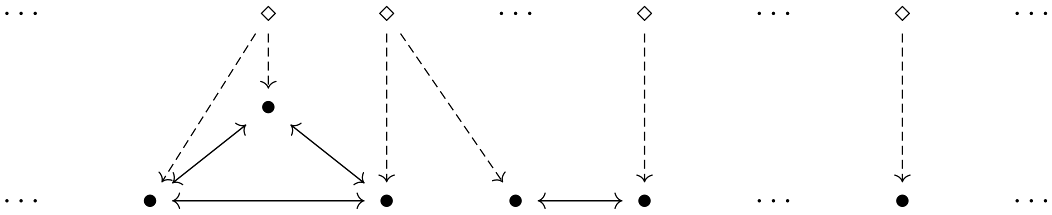

Figure 3.2: Network sampling of siblings ($\bullet$) via households ($\diamond$)

Notice that the OP of network sampling requires the siblings to report each other, such that it is network exhaustive if at least one sibling in a network belongs to an initial sample of households, no matter how many siblings there are otherwise. For this reason, we let each network of siblings form a clique in G.

Let Ω consist of all the sibling networks. Let F be the initial sampling frame (consisting of the household $\diamond$-nodes) for graph sampling from G in Figure 3.2. As shown in Figure 3.3, the strategy BIGS-IWE is applicable for network sampling from G, with respect to

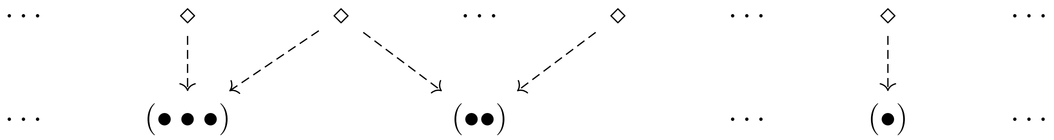

Figure 3.3: BIG constructed by (3.1) for network sampling in Figure 3.2.

$\mathcal{B} = (F, \Omega; H)$, where each distinct network in G is represented by an element κ in Ω, and $(i\kappa) \in H$ for any $i \in F$ and $\kappa \in \Omega$ iff (3.1). Condition (i) of Theorem 3.1 is satisfied, if every network has a positive sample inclusion probability under network sampling, whether or not it is considered as a graph sampling method. Condition (ii) is satisfied if the multiplicity of each $\kappa \in \Omega_s$ is observed under network sampling from G. This requires one to identify the household of each sibling in a sampled network.

3.3 Line-intercept sampling

Line-intercept sampling (LIS) is a method of sampling habitats in a given area, where a habitat is sampled if a chosen line segment transects it. The habitats are of irregular shapes, such as animal tracks or forests. In a simpler setting, each transect line is selected at random by selecting randomly a position along a fixed *baseline* that traverses the whole study area, in the direction perpendicular to the baseline. In a more general setting, a point is randomly selected on the map and an angle is randomly chosen, yielding a line segment of fixed length or transecting the whole area in the chosen direction. Repetition of either procedure generates an IID sample of lines.

3.3.1 Constructed BIGS

Becker (1991) gives an example of baseline-LIS. The aim is to estimate the total number of wolverines, denoted by θ, in the boxed area sketched in Figure 3.4.

Four systematic samples A, B, C and D, each containing 3 positions, are drawn on the baseline that is equally divided into 3 sections of length 12 miles each. Following the 12 selected lines *and* any wolverine track that intercepts them yields the four observed tracks, denoted by $\kappa = 1, ..., 4$ and heuristically indicated by the dashed lines in Figure 3.4. Let y_κ be the associated number of wolverines, and L_κ the length of the projection of κ on the baseline. From top to bottom and left to right, we observe $(y_1, L_1) = (1, 5.25)$, $(y_2, L_2) = (2, 7.5)$, $(y_3, L_3) = (2, 2.4)$ and $(y_4, L_4) = (1, 7.05)$.

Given the observed tracks, partition the baseline into *projection segments*, each with associated length x_i, for $i = 1, ..., m_F^*$, where $m_F^* = 7$. The probability that the i-th projection segment is selected under systematic sampling here is $p_i = x_i/12$ where, from left to right, x_1 refers to the overlapping projection of $\kappa = 1$ and 2, x_2 the projection of $\kappa = 2$ that does not overlap with $\kappa = 1$, x_3 the distance between projections of $\kappa = 2$ and 3, x_4 the projection of $\kappa = 3$, x_5 the distance between projections of $\kappa = 3$ and 4, x_6 the projection of $\kappa = 4$, and x_7 the distance between $\kappa = 4$ and right-hand border.

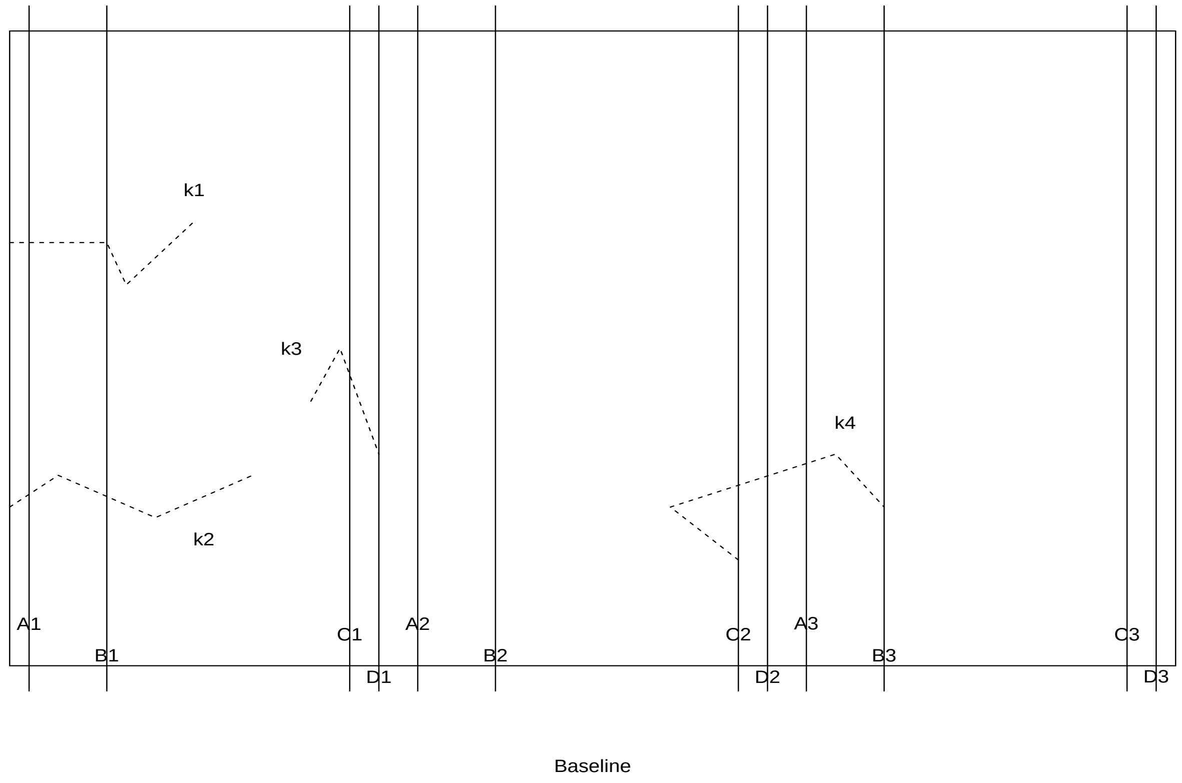

Figure 3.4: LIS with 4 systematic samples (A, B, C, D) of 3 positions each.

On the r-th draw, let s_r contain the selected projection segments, and Ω_r the wolverine tracks that intercept the sampled line originating from each $i \in s_r$. In this particular case, we have $s_1 = s_2 = \{1, 5, 6\}$, yielding $\Omega_1 = \Omega_2 = \{1, 2, 4\}$ on the first two draws A and B, and $s_3 = s_4 = \{4, 6, 7\}$, yielding $\Omega_3 = \Omega_4 = \{3, 4\}$ on the last two draws C and D. The distinct projection segments selected over all the draws are $s = \bigcup_{r=1}^{4} s_r = \{1, 4, 5, 6, 7\}$, and the distinct tracks are $\Omega_s = \bigcup_{r=1}^{4} \Omega_r = \{1, 2, 3, 4\}$.

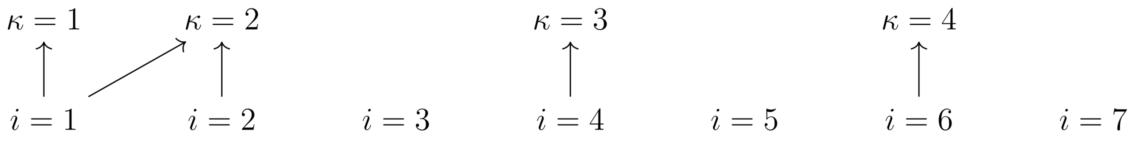

Figure 3.5: BIG constructed from realised line-intercept samples

Let $F^* = \{i : i = 1, ..., m_F^*\}$ contain the projection segments constructed from the actual samples s and Ω_s. Let $\mathcal{B}^* = (F^*, \Omega_s; H^*)$ be given by Figure 3.5, where H^* consists of the incident observational links from F^* to Ω_s satisfying (3.1). Denote by β_κ^* the ancestors of κ in $\mathcal{B}^*$, where $\beta_1^* = \{1\}$, $\beta_2^* = \{1, 2\}$, $\beta_3^* = \{4\}$ and $\beta_4^* = \{6\}$.

Let $\Omega = \{1, ..., \kappa, ..., |\Omega|\}$ contain all the wolverine tracks in the area, where $|\Omega| \geq 4$. Let $F = \{1, ..., i, ..., m_F\}$ be the sampling frame, which consists of all the projection segments that can be constructed from Ω. Let $H = \{(i\kappa); i \in F, \kappa \in \Omega\}$ by (3.1), where an edge exists from i to κ provided κ intercepts any line that originates from the i-th projection segment. The population BIG is given by $\mathcal{B} = (F, \Omega; H)$.

In practice, only $\mathcal{B}^*$ can be constructed but not $\mathcal{B}$. The two are not the same generally, in that one needs to further partition the projection segments of F^* in F based on Ω, in order to accommodate the unobserved tracks in $\Omega \setminus \Omega_s$. For instance, suppose there is a track that can only be intercepted from the 7-th projection segment in F^* and the track does not reach the right-hand border, then this projection segment would be partitioned into 3 segments in F, and (F, H) would differ from (F^*, H^*) accordingly.

Under LIS, field observation along a line has an actual width of detectability. Dividing the baseline accordingly yields thus a known sampling frame F' of detectability partitions. Let $\mathcal{B}' = (F', \Omega; H')$ be given by (3.1). Clearly, both the conditions (i) and (ii) of Theorem 3.1 are satisfied by the OP of LIS, so that the strategy BIGS-IWE can be applied with respect to $\mathcal{B}'$ to yield unbiased estimation of θ.

Now, as along as the unit of detectability is negligible in scale compared to the baseline, one can assume the elements of F' to be nested in those of F^* (or F), such that the selection probability of each observed track κ with respect to BIGS from $\mathcal{B}'$ can be correctly calculated using $\mathcal{B}^*$ (or $\mathcal{B}$). Thus, the strategy BIGS-IWE defined for $\mathcal{B}'$ can be applied using the observed $\mathcal{B}^*$, just as when $\mathcal{B}$ were known.

3.3.2 HTE

When the position of the transect line is selected randomly over the entire baseline, the selection probability of track κ on each draw is given by $p_{(\kappa)} = \sum_{i \in \beta^*_\kappa} p_i$, calculated with respect to $\mathcal{B}^*$, where $p_i = x_i/12$. The inclusion probability of $\kappa \in \Omega_s$ is one minus the probability that κ is not selected on any of the 4 draws, $\pi_{(\kappa)} = 1 - (1 - p_{(\kappa)})^4$. Denote by $p_{(\kappa \cup \ell)} = \sum_{i \in \beta^*_\kappa \cup \beta^*_\ell} p_i$ the probability of selecting either κ or ℓ on a given draw. The second-order inclusion probability of $\kappa \neq \ell \in \Omega_s$ is given by

$$\begin{aligned}\pi_{(\kappa\ell)} &= 1 - \Big(\Pr(\kappa \notin \Omega_s) + \Pr(\ell \notin \Omega_s) - \Pr(\kappa \notin \Omega_s, \ell \notin \Omega_s)\Big) \\ &= \pi_{(\kappa)} + \pi_{(\ell)} - 1 + \big(1 - p_{(\kappa\cup\ell)}\big)^4\end{aligned}$$

The computation of $p_{(\kappa\cup\ell)}$ involves some extra details when a systematic sample of 3 positions are drawn each time, which can be found in Thompson (2012). We have

$$p_{(1)} = 0.4375 \qquad p_{(2)} = 0.625 \qquad p_{(3)} = 0.2 \qquad p_{(4)} = 0.5875$$
$$\pi_{(1)} = 0.90 \qquad \pi_{(2)} = 0.98 \qquad \pi_{(3)} = 0.59 \qquad \pi_{(4)} = 0.97$$
$$\pi_{(12)} = 0.90 \quad \pi_{(13)} = 0.51 \quad \pi_{(14)} = 0.88 \quad \pi_{(23)} = 0.57 \quad \pi_{(24)} = 0.95 \quad \pi_{(34)} = 0.59$$

Accordingly, the HTE and its estimated variance are

$$\hat{\theta}_y = \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}} = 7.57 \qquad \text{and} \qquad \hat{V}(\hat{\theta}_y) = 5.27$$

3.3.3 HH-type estimators

An unbiased estimator of θ from the r-th draw is

$$\tau_r = \sum_{\kappa \in \Omega_r} \frac{y_\kappa}{p_{(\kappa)}}$$

where $\tau_1 = \tau_2 = 7.1878$ and $\tau_3 = \tau_4 = 11.7021$. The Hansen-Hurwitz (HH) estimator over all the draws (with replacement) and its estimated variance are

$$\hat{\theta}_{HH} = \frac{1}{4}\sum_{r=1}^{4} \tau_r = 9.44 \qquad \text{and} \qquad \hat{V}(\hat{\theta}_{HH}) = 1.70$$

By the strategy BIGS-IWE using $\mathcal{B}^*$, the HH-type estimator (2.7) on the r-th draw is

$$\hat{\theta}_{z,r} = \sum_{i \in s_r} \frac{z_i}{p_i} \qquad \text{where} \qquad z_i = \sum_{\kappa \in \alpha_i^*} \omega_{i\kappa} y_\kappa$$

The multiplicity estimator $\hat{\theta}_{z\beta}$ uses equal weights $\omega_{i\kappa} = 1/|\beta_\kappa|$, where $\omega_{11} = \omega_{43} = \omega_{64} = 1$, and $\omega_{12} = \omega_{22} = 0.5$. The resulting IWE and its estimated variance are given by

$$\hat{\theta}_{z\beta,1} = \hat{\theta}_{z\beta,2} = 6.2736 \qquad \hat{\theta}_{z\beta,3} = \hat{\theta}_{z\beta,4} = 11.7021$$
$$\hat{\theta}_{z\beta} = 8.99 \qquad \text{and} \qquad \hat{V}(\hat{\theta}_{z\beta}) = 2.46$$

The PIDA weights (2.9) with $\gamma = 0$ yield $\omega_{i\kappa} = p_i/p_{(\kappa)}$ for $i \in \beta_\kappa$, where $\omega_{11} = \omega_{43} = \omega_{64} = 1$, $\omega_{12} = p_1/p_{(2)} = 0.7$ and $\omega_{22} = p_2/p_{(2)} = 0.3$. This yields

$$\hat{\theta}_{z,r} = \sum_{i \in s_r} \frac{1}{p_i} \sum_{\kappa \in \alpha_i} \frac{p_i}{p_{(\kappa)}} y_\kappa = \sum_{\kappa \in \Omega_r} \frac{y_\kappa}{p_{(\kappa)}} = \tau_r$$

on the r-th draw. Thus, the HH-estimator is in fact the IWE $\hat{\theta}_{z\alpha 0}$. One can also use other values of γ in (2.9). For instance, given $\gamma = 0.5$, we have $\omega_{11} = \omega_{43} = \omega_{64} = 1$, $\omega_{12} = 0.6226$ and $\omega_{22} = 0.3773$. The resulting IWE $\hat{\theta}_{z\alpha.5}$ and its estimated variance are given by

$$\hat{\theta}_{z\alpha.5,1} = \hat{\theta}_{z\alpha.5,2} = 6.8341 \qquad \hat{\theta}_{z\alpha.5,3} = \hat{\theta}_{z\alpha.5,4} = 11.7021$$
$$\hat{\theta}_{z\alpha.5} = 9.27 \qquad \text{and} \qquad \hat{V}(\hat{\theta}_{z\alpha.5}) = 1.97$$

3.3.4 Discussion

The results for all the IWEs above are summarised in Table 3.1, where $\hat{\theta}_{HH}$ is equal to $\hat{\theta}_{z\alpha 0}$. Neither the HTE $\hat{\theta}_y$ nor the multiplicity estimator $\hat{\theta}_{z\beta}$ is efficient here. Efficiency gains can be achieved using the PIDA weights (2.9). In this case, adjusting the equal weights by the selection probability while disregarding the degrees of the nodes in F performs well, where $\hat{\theta}_{z\alpha 0}$ has the lowest estimated variance. Notice that since the variances in Table 3.1 are estimates, the true variance of $\hat{\theta}_{z\alpha 0}$ may or may not be smaller than some $\hat{\theta}_{z\alpha\gamma}$

with $\gamma \neq 0$. Meanwhile, setting $\gamma = 1.227$ would numerically reproduce the equal weights $\omega_{12} = \omega_{22} = 0.5$ based on the observed sample. It seems that the IWE by (2.9) has the potential to approximate the relatively more efficient estimators in different situations, if one is able to choose the coefficient γ in (2.9) appropriately.

Table 3.1: Some BIGS-IWE using $\mathcal{B}^*$.

	$\hat{\theta}_y$	$\hat{\theta}_{z\beta}$	$\hat{\theta}_{z\alpha 0}$	$\hat{\theta}_{z\alpha .5}$
Estimate of θ	7.57	8.99	9.44	9.27
Variance Estimate	5.27	2.46	1.70	1.97

Given the systematic sampling design of the transect lines, the tracks $\{1, 2, 4\}$ can only be observed if a position is selected in the left part of 1st projection segment, which would only result in $\{1, 5, 6\}$ as the sampled projection segments. Similarly, the tracks $\{3, 4\}$ can only be observed if a position is selected in 4th projection segment, which would only result in $\{4, 6, 7\}$ as the sampled projection segments. Thus, applying the RB method would not change any unbiased IWE given the observed sample Ω_s in this case.

3.4 Sampling from relational databases

Let $G = (U, A)$ be the structure of a relational database, where U contains the keys of all the tables. Suppose three types of entities: stock transaction $\{h, i, j, ...\}$, buyer $\{b, c, ...\}$, and company $\{\kappa, \ell, ...\}$. Let $(hb) \in A$ if transaction h involves buyer b, where the mapping from transactions to buyers is many-to-one. Let $(b\kappa) \in A$ if buyer b has ownership in company κ, where the mapping from buyers to companies is many-to-many.

3.4.1 BIGS

Given an initial sample s_0 from all the transactions F, suppose one traces first their buyers and then the companies in which these buyers have ownership. Let Ω consist of the companies that can be traced in this way. Let $\theta = \sum_{\kappa \in \Omega} y_\kappa$. Suppose the y-values are unknown to start with, but can be collected for a given sample of companies at some cost, in order to estimate θ. Suppose it is possible to identify the ancestors in F of any $\kappa \in \Omega_s$ by cross-checking the database tables.

Let Q contain all the buyers. Some configurations of G are given as (I) - (III) in Figure 3.6, where $\mathcal{B}_1 = (F, Q; H_1)$ and $\mathcal{B}_2 = (Q, \Omega; H_2)$ are fused together by Q. Using configuration (II) to illustrate the notation, we have

$$\begin{cases} \alpha_h = \{b\},\ \alpha_h^2 = \alpha(\alpha_h) = \{\kappa\} \\ \alpha_i = \{c\},\ \alpha_i^2 = \alpha(\alpha_i) = \{\kappa, \ell\} \\ \alpha_j = \{c\},\ \alpha_j^2 = \alpha(\alpha_j) = \{\kappa, \ell\} \end{cases} \qquad \begin{cases} \beta_\kappa = \{b, c\},\ \beta_\kappa^2 = \beta(\beta_\kappa) = \{h, i, j\} \\ \beta_\ell = \{c\},\ \beta_\ell^2 = \beta(\beta_\ell) = \{i, j\} \end{cases}$$

Denote by $\mathcal{B}^* = (F, \Omega; H^*)$ the BIG constructed by (3.1), which only consists of the links from transactions to companies, where $(i\kappa) \in H^*$ iff $\kappa \in \alpha_i^2$ in G. This gives rise to

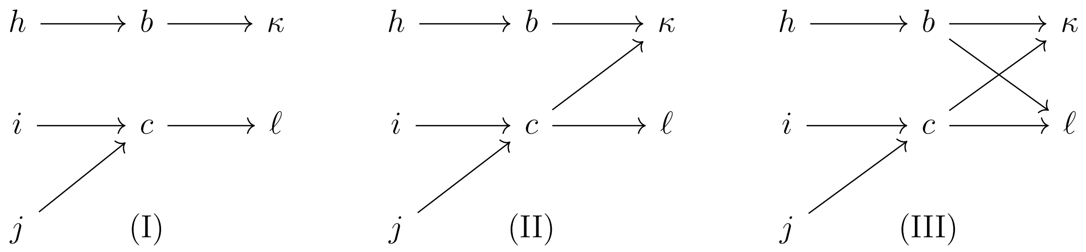

Figure 3.6: Some configurations of relational database structure

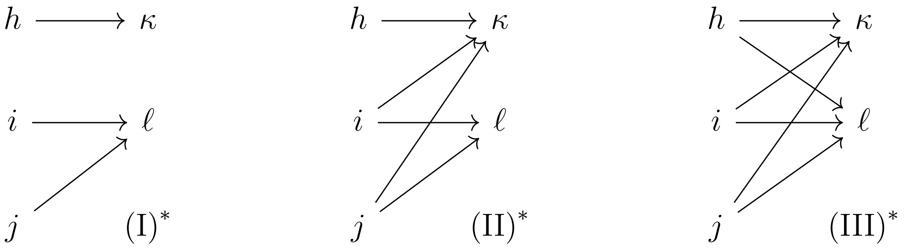

Figure 3.7: BIG configurations corresponding to Figure 3.6

(I)* - (III)* in Figure 3.7 instead of (I) - (III), respectively.

3.4.2 IWE

The HTE (2.6) is based on Ω_s and $\pi_{(\kappa)}$ for $\kappa \in \Omega_s$, which are the same either calculated with respect to G (as in Figure 3.6) or $\mathcal{B}^*$ (as in Figure 3.7).

Under BIGS from $\mathcal{B}^*$, one can directly apply another IWE based on H^*. To emphasise the dependence on $\mathcal{B}^*$, the HH-type estimator (2.7) can be given as

$$\hat{\theta}_z^* = \sum_{i \in s_0} \frac{z_i^*}{\pi_i} \qquad \text{and} \qquad z_i^* = \sum_{\kappa \in \alpha_i^*} \omega_{i\kappa}^* y_\kappa \tag{3.2}$$

where α_i^* and β_κ^* refer to $\mathcal{B}^*$, and $\sum_{i \in \beta_\kappa^*} \omega_{i\kappa}^* = 1$ for any $\kappa \in \Omega$.

Meanwhile, there are different ways of using fixed incident weights defined for $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively, where ω_{ib} for $i \in F$ and $b \in \alpha_i \subset Q$ and $\sum_{i \in \beta_b} \omega_{ib} = 1$ for each $b \in Q$, and $\omega_{b\kappa}$ for $b \in Q$ and $\kappa \in \alpha_b \subseteq \Omega$ and $\sum_{b \in \beta_\kappa} \omega_{b\kappa} = 1$ for each $\kappa \in \Omega$.

Let $\delta_b = 1$ if $b \in Q$ is sampled and 0 otherwise. Let the *HT-HH estimator* be given by

$$\hat{\theta}_{yz} = \sum_{b \in Q} \frac{\delta_b}{\pi_{(b)}} z_b \qquad \text{and} \qquad z_b = \sum_{\kappa \in \alpha_b} \omega_{b\kappa} y_\kappa \tag{3.3}$$

That is, a value z_b is constructed for $b \in Q$ using the HH-type weights between Q and Ω, which is then used in the manner of HTE based on the links between F and Q. The estimator (3.3) is unbiased, because

$$E(\hat{\theta}_{yz}) = \sum_{b \in Q} z_b = \sum_{\kappa \in \Omega} y_\kappa$$

Let $\omega_{i\kappa} = \sum_{b\in\alpha_i} \omega_{ib}\omega_{b\kappa}$ for $i \in F$ and $\kappa \in \alpha_i^2$. Let the *HH-HH estimator* be given by

$$\hat{\theta}_{zz} = \sum_{i\in s_0} \frac{z_i}{\pi_i} \qquad \text{and} \qquad z_i = \sum_{\kappa\in\alpha_i^2} \omega_{i\kappa} y_\kappa = \sum_{b\in\alpha_i} \sum_{\kappa\in\alpha_b} \omega_{ib}\omega_{b\kappa} y_\kappa \tag{3.4}$$

That is, y_κ is apportioned to $z_{b\kappa} = \omega_{b\kappa} y_\kappa$ using the intermediary nodes b, and each $z_{b\kappa}$ is apportioned to z_i according to the weights ω_{ib}. Thus, y_κ is apportioned to z_i by the compound weights $\omega_{i\kappa}$. The estimator (3.4) is unbiased, because

$$E(\hat{\theta}_{zz}) = \sum_{i\in F} z_i = \sum_{i\in F} \sum_{b\in\alpha_i} \sum_{\kappa\in\alpha_b} \omega_{ib}\omega_{b\kappa} y_\kappa = \sum_{\kappa\in\Omega} y_\kappa \sum_{b\in\beta_\kappa} w_{b\kappa} \sum_{i\in\beta_b} \omega_{ib} = \sum_{\kappa\in\Omega} y_\kappa$$

3.4.3 A special case

The estimators (3.2), (3.3) and (3.4) are different ways of applying the strategy BIGS-IWE in the present situation.

In the special case of $|s_0| = 1$ and $|\alpha_i| \equiv 1$ for all $i \in F$, they can be made identical by applying the PIDA weights (2.9) with $\gamma = 0$ directly everywhere, yielding

$$\omega_{i\kappa}^* = \frac{\pi_i}{\sum_{j\in\beta_\kappa^2} \pi_j} \qquad \text{and} \qquad \omega_{ib} = \frac{\pi_i}{\sum_{j\in\beta_b} \pi_j} \qquad \text{and} \qquad \omega_{b\kappa} = \frac{\pi_{(b)}}{\sum_{c\in\beta_\kappa} \pi_{(c)}}$$

where the weights sum to 1 over the corresponding ancestor set. We have

$$\pi_{(b)} = \sum_{j\in\beta_b} \pi_j \qquad \text{and} \qquad \sum_{c\in\beta_\kappa} \pi_{(c)} = \sum_{c\in\beta_\kappa} \sum_{j\in\beta_c} \pi_j$$

now that $|s_0| = 1$. Since $|\alpha_i| \equiv 1$, we have $\beta_c \cap \beta_b = \emptyset$ for any $b \neq c \in Q$, such that

$$\sum_{c\in\beta_\kappa} \sum_{j\in\beta_c} \pi_j = \sum_{j\in\beta_\kappa^2} \pi_j$$

It follows that $w_{i\kappa}^* = \omega_{ib}\omega_{b\kappa}$ and $\hat{\theta}_z^* = \hat{\theta}_{zz}$. Since $a_{ib_i} = 1$ for only one $b_i \in Q$, we have

$$\begin{aligned}\hat{\theta}_{yz} &= \sum_{i\in s_0} \sum_{b\in\alpha_i} \frac{z_{(b)}}{\pi_{(b)}} = \sum_{i\in s_0} \frac{1}{\pi_{(b_i)}} \sum_{\kappa\in\alpha_{b_i}} \omega_{b_i\kappa} y_\kappa = \sum_{i\in s_0} \frac{1}{\pi_{(b_i)}} \sum_{\kappa\in\alpha_{b_i}} \frac{\pi_{(b_i)}}{\sum_{j\in\beta_\kappa^2} \pi_j} y_\kappa \\ &= \sum_{i\in s_0} \sum_{\kappa\in\alpha_i^2} \frac{y_\kappa}{\sum_{j\in\beta_\kappa^2} \pi_j} = \sum_{i\in s_0} \sum_{\kappa\in\alpha_i^2} \omega_{i\kappa}^* \frac{y_\kappa}{\pi_i} = \hat{\theta}_z^*\end{aligned}$$

To illustrate, take (II) and (II)* as the population graphs. Let $y_\kappa \equiv 1$. We have

$$\begin{cases} \omega_{b\kappa} = \pi_h \\ \omega_{c\kappa} = \pi_i + \pi_j \\ \omega_{c\ell} = 1 \end{cases} \text{and} \begin{cases} \omega_{hb} = 1 \\ \omega_{ic} = \frac{\pi_i}{\pi_i+\pi_j} \\ \omega_{jc} = \frac{\pi_j}{\pi_i+\pi_j} \end{cases} \Rightarrow \begin{cases} \omega^*_{h\kappa} = \pi_h \\ \omega^*_{i\kappa} = \pi_i \\ \omega^*_{j\kappa} = \pi_j \end{cases} \text{and} \begin{cases} \omega^*_{i\ell} = \frac{\pi_i}{\pi_i+\pi_j} \\ \omega^*_{j\ell} = \frac{\pi_j}{\pi_i+\pi_j} \end{cases}$$

$$\begin{aligned} \hat{\theta}^*_z &= \frac{\delta_h}{\pi_h}\pi_h + \frac{\delta_i}{\pi_i}\big(\pi_i + \frac{\pi_i}{\pi_i+\pi_j}\big) + \frac{\delta_j}{\pi_j}\big(\pi_j + \frac{\pi_j}{\pi_i+\pi_j}\big) \\ &= \delta_h + \delta_i\big(1 + \frac{1}{\pi_i+\pi_j}\big) + \delta_j\big(1 + \frac{1}{\pi_i+\pi_j}\big) \\ V(\hat{\theta}^*_z) &= E(\hat{\theta}^{*2}_z) - E(\hat{\theta}^*_z)^2 = \pi_h + \pi_i\big(1 + \frac{1}{\pi_i+\pi_j}\big)^2 + \pi_j\big(1 + \frac{1}{\pi_i+\pi_j}\big)^2 - 4 \\ &= 3 + \frac{1}{\pi_i+\pi_j} - 4 = \frac{1}{\pi_i+\pi_j} - 1 = \frac{\pi_h}{\pi_i+\pi_j} \\ \hat{\theta}_{yz} &= \frac{\delta_b}{\pi_h}\pi_h + \frac{\delta_c}{\pi_i+\pi_j}(\pi_i+\pi_j+1) = \delta_b + \delta_c\big(1 + \frac{1}{\pi_i+\pi_j}\big) = \hat{\theta}^*_z \end{aligned}$$

since $\delta_c = \delta_i + \delta_j$ here. Indeed, the HTE is the same here as well, since

$$\hat{\theta}_y = 1 + \frac{\delta_\ell}{\pi_i+\pi_j} = (\delta_h + \delta_i + \delta_j) + \frac{\delta_i + \delta_j}{\pi_i+\pi_j} = \hat{\theta}^*_z$$

To see that the PIDA weighting scheme (2.9) matters, consider (III), where

$$\begin{cases} \omega_{b\kappa} = \omega_{b\ell} = \pi_h \\ \omega_{c\kappa} = \omega_{c\ell} = \pi_i + \pi_j \end{cases} \quad \text{and} \quad \begin{cases} \omega_{hb} = 1 \\ \omega_{ic} = \frac{\pi_i}{\pi_i+\pi_j} \\ \omega_{jc} = \frac{\pi_j}{\pi_i+\pi_j} \end{cases} \Rightarrow \begin{cases} z_b = 2\pi_h \\ z_c = 2(\pi_i+\pi_j) \end{cases}$$

$$\hat{\theta}_{yz} = \frac{\delta_b}{\pi_h}2\pi_h + \frac{\delta_c}{\pi_i+\pi_j}2(\pi_i+\pi_j) \equiv 2 \quad \text{and} \quad V(\hat{\theta}_{yz}) = 0$$

whilst $V(\hat{\theta}_{yz}) > 0$ using the multiplicity weights between Q and Ω in (III), since

$$\omega_{b\kappa} = \omega_{c\kappa} = \omega_{c\ell} = 0.5 \quad \Rightarrow \quad z_b = z_c = 1 \quad \Rightarrow \quad \hat{\theta}_{yz} = \frac{\delta_b}{\pi_h} + \frac{\delta_c}{\pi_i+\pi_j} \ .$$

Bibliographic notes

Birnbaum and Sirken (1965) pioneered the unbiased estimators under multiplicity, indirect or network sampling. Network sampling is e.g. considered by Sirken (1970), Sirken and Levy (1974), Sirken (2004, 2005). See Lavalleè (2007) for a treatise of indirect sampling. Frank (1977c) notes that graph representation is possible for such methods, which "are not explicitly stated as graph problems but which can be given such formulations".

Zhang and Oguz-Alper (2020) establish essentially Theorem 3.1, which specifies the conditions for applying the BIGS-IWE strategy to an arbitrary graph sampling situation. The formulation of the theorem is revised here to avoid potential confusions, based on feedbacks received via private communications.

Kaiser (1983) considers the general setting of LIS, where a point is randomly selected on the map and an angle is randomly chosen, yielding a line segment of fixed length or transecting the whole area in the chosen direction. The application of BIGS-IWE described by Patone and Zhang (2020) for the simpler setting of baseline-LIS can be extended to the general setting. Becker (1991) applies the HH-estimator, which is shown to be a special case of the IWE. Thompson (2012, Ch. 19.1) describes the HTE under LIS, without using the BIGS set-up. The HTE is another special case of the IWE. As illustrated in Section 3.3, infinitely many unbiased incident weighting estimators can be considered when applying the BIGS-IWE strategy to LIS.

All these unconventional sampling techniques are characterised by the presence of some deterministic rules of observation, *in addition* to the initial sampling design. The application of the BIGS-IWE strategy clarifies the common structure underlying these seemingly unrelated problems, which consists of the distinction between sampling units (household or line) and study units (sibling network or wild-life habitat), as well as the many-to-many observational links between them.

In particular, the information of multiplicity must be collected *in addition to* the sample of study units. For sampling of sibling networks via households, this refers to the observational links between each sampled network and the initial frame of households. For sampling of habitats via line segments, one needs to identify the other line segments that would have led to each sampled habitat. The requirement of such additional information coincide formally with that of the ancestry knowledge under BIGS.

Provided the BIGS-IWE strategy is applicable, one can employ the general class of IWE, which includes as special cases the specific estimators that are already known for the different problems, such as the HH-type estimator for network sampling and the HTE for LIS. Moreover, other unbiased IWEs may be developed in future. Such generality implies potentials of efficiency gain.

Chapter 4

Adaptive sampling

An *adaptive* OP depends on the values associated with the nodes or edges, in addition to the graph itself. *Adaptive network sampling (ANS)* is the case if the adaptive OP is network exhaustive. The graph G may or may not be known; the values needed for the adaptive OP are unknown generally.

4.1 ACS from known graphs

Given an initial probability sample of units, *adaptive cluster sampling (ACS)* refers to sampling designs in which, whenever the variable of interest of a selected unit satisfies a given criterion, additional units in the neighbourhood of that unit are added to the sample. ACS is a special case of ANS, as long as "neighbourhood" is defined in terms of the connections among the units, whether or not ACS is given as a graph problem.

4.1.1 Spatial ACS

Consider the example of ACS in 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 (contiguous grids) which are adjacent in the undirected simple graph $G = (U, A)$ in Figure 4.1, where we simply denote each grid by its y-value and refer to it as a node below. The graph G is known, but the y-values associated with the nodes are unknown.

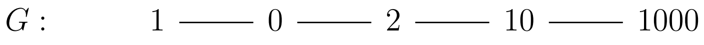

Figure 4.1: A spatial graph for ACS (Thompson, 1990)

Given an initial sample s_0 of size 2 by SRS from $F = U$, one would survey their y-values, and all the adjacent nodes of a sample node i only if $y_i \geq 5$. The procedure is repeated for all the acquired nodes, which may or may not generate additional nodes to be surveyed. The process is terminated, when the last observed nodes are all below the threshold. The interest is to estimate the mean of y over U, denoted by μ. In terms of

(1.1), we have

$$\Omega = U \qquad \text{and} \qquad \theta = \sum_{i \in \Omega} y_i \qquad \text{and} \qquad \mu = \theta / N$$

A network in this spatial graph may consist of a single node, whose y-value is below the threshold, to be referred to as a *terminal node* because ACS would not proceed further from such a node; or it may consist of a cluster of nodes, where $y_i \geq 5$ for every node i in the network, to be referred to as a *non-terminal (NT) network*. An edge node is a terminal node that is adjacent to one or several NT networks, such as 2 to the NT network $\{10, 1000\}$ in Figure 4.1, where 2 can be observed from 10 or 1000 but 10 or 1000 cannot be observed from 2. The inclusion probability of an edge node under ACS from G cannot be calculated correctly unless s_0 is guaranteed always to intersect *all* its NT networks, because the ancestry knowledge would be lacking otherwise.

The adaptive OP of ACS is network exhaustive in G, since a whole network is sampled if at least one of its nodes is sampled.

4.1.2 Two strategies using BIGS

ACS from G can be given as BIGS from $\mathcal{B} = (F, \Omega; H)$ with $F = U$ in Figure 4.2. The same grid is referred to as κ in Ω and i_κ in F. This BIG contains all the observational links under ACS from G. However, the OP of ACS does not ensure the ancestry knowledge of an edge node in $\mathcal{B}$, for which the condition (ii) of Theorem 3.1 is violated.

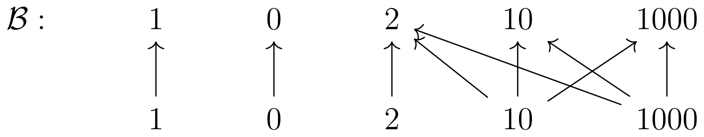

Figure 4.2: BIG by (3.1) for ACS from G in Figure 4.1.

For a feasible strategy, one may use a *modified HTE*, where an edge node is *eligible* for estimation, *only if* it is selected in s_0 directly. This probability is known, denoted by $\phi_\kappa = \Pr(i_\kappa \in s_0)$ for each $\kappa \in \Omega$. That is, edge or non-edge, terminal nodes are eligible for estimation in the same way. Let the modified HTE be

$$\hat{\theta}^*_{HT} = \sum_{\kappa \in \Omega_s} W_\kappa y_\kappa = \sum_{\kappa \in \Omega} \mathbb{I}(\kappa \in \Omega_s) W_\kappa y_\kappa$$

where $W_\kappa = \pi^{-1}_{(\kappa)}$ for any non-terminal node, and

$$W_\kappa = \begin{cases} \phi_\kappa^{-1} & \text{if } i_\kappa \in s_0 \\ 0 & \text{otherwise} \end{cases}$$

for any terminal node κ. The estimator $\hat{\theta}^*_{HT}$ is unbiased because, for any $\kappa \in \Omega$, we have

$$E\big(\mathbb{I}(\kappa \in \Omega_s) W_\kappa\big) = 1$$

Denote this strategy by ($\mathcal{B}$, $\hat{\theta}^*_{HT}$), which is not one of BIGS-IWE, because the condition (ii) of Theorem 3.1 is not satisfied and the IWE (2.2) is not applicable to $\mathcal{B}$.

For a genuine BIGS-IWE strategy, consider $\mathcal{B}^* = (F, \Omega; H^*)$ in Figure 4.3. As a modification of β_κ by (3.1), the observational links $(10, 2)$ and $(1000, 2)$ in $\mathcal{B}$ are removed, such that $\beta^*_2 = \{2\}$ under BIGS from $\mathcal{B}^*$ instead of $\beta_2 = \{2, 10, 1000\}$ under BIGS from $\mathcal{B}$. The OP of ACS ensures the ancestry knowledge in $\mathcal{B}^*$, since obviously β^*_2 is always identified whenever $2 \in \Omega_s$, e.g. given $s_0 = \{0, 2\}$ or $\{2, 10\}$. One can now use the (unmodified) HTE (2.6) directly. Denote this strategy by $(\mathcal{B}^*, \hat{\theta}_y)$.

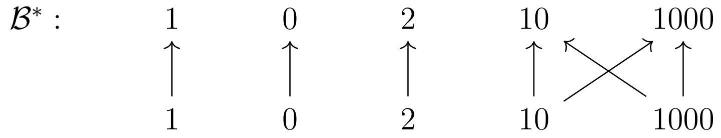

Figure 4.3: BIG by modifying (3.1) for ACS from G in Figure 4.1.

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

It is possible to apply other IWEs with respect to BIGS from $\mathcal{B}^*$. Now that the nodes of a network are all observed together under ACS if any of them is observed, $|\alpha_i|$ in the PIDA weight (2.9) is observable. However, since $|\alpha_i|$ is the same for the nodes in the same network and the initial sampling is SRS, the PIDA weights are all equal, so that the estimator $\hat{\theta}_{z\alpha\gamma}$ reduces to the multiplicity estimator $\hat{\theta}_{z\beta}$.

4.1.3 A sample-dependent strategy using BIGS

As another possible modification of β_κ by (3.1), one can make the edge node 2 eligible, *only if* s_0 intersects its NT network $\{10, 1000\}$, whether or not $2 \in s_0$. Figure 4.4 gives the corresponding $\mathcal{B}^\dagger = (F, \Omega; H^\dagger)$, where $\beta^\dagger_2 = \{10, 1000\}$. The HTE can be defined with respect to BIGS from $\mathcal{B}^\dagger$, yielding the strategy $(\mathcal{B}^\dagger, \hat{\theta}_y)$. Note that the OP of ACS ensures that $\beta^\dagger_2$ is entirely observed whenever $s_0 \cap \beta^\dagger_2 \neq \emptyset$.

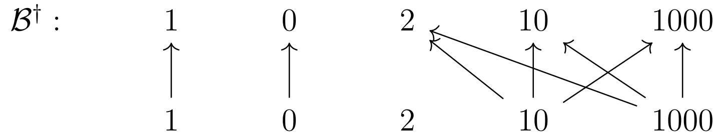

Figure 4.4: Another BIG by modifying (3.1) for ACS from G in Figure 4.1.

Whereas the strategies $(\mathcal{B}, \hat{\theta}^*_y)$ and $(\mathcal{B}^*, \hat{\theta}_y)$ can be determined *a priori*, the strategy $(\mathcal{B}^\dagger, \hat{\theta}_y)$ is *sample-dependent* in the sense that it can only be identified when the NT network $\{10, 1000\}$ intersects s_0, the probability of which is $1 - \bar{\pi}_{10,1000} = 0.7$ under ACS from Figure 4.1. Meanwhile, there is a probability of 0.2 that 2 is observed on its own while neither 10 nor 1000 is selected in s_0, in which case one needs to adopt $\mathcal{B}^*$ instead.

In other words, for a sample-dependent BIGS-IWE strategy for ACS from G, one can modify β_κ by (3.1) for a terminal node i as follows:

- use $\beta_i^* = \{i\}$ if i is observed on its own (without any adjacent NT network),
- use an adjacent NT network $\beta_i^\dagger$ if i is observed together with $\beta_i^\dagger$.

Thus, whenever 2 is observed by ACS from Figure 4.1, the strategy $(\mathcal{B}^*, \hat{\theta}_y)$ is adopted 2 out 9 times, whereas $(\mathcal{B}^\dagger, \hat{\theta}_y)$ is adopted the rest of times. Since only one sample graph is realised in reality, inference needs to be conditional on the chosen strategy.

4.1.4 Strategy BIGS-IWE generally

A general approach for applying the strategy BIGS-IWE to graph sampling from G can be given as follows. Let $\mathcal{B}$ be the BIG, where β_κ contains all the ancestors of κ satisfying (3.1). Suppose the strategy using $\mathcal{B}$ is not directly applicable because the condition (ii) of Theorem 3.1 is not satisfied by the OP in G. For any $\kappa \in \Omega_s$, let $\beta_{\kappa|s}$ be the observed *(sample) ancestors* satisfying (3.1) based on the sample graph G_s, where

$$\emptyset \neq \beta_{\kappa|s} \subseteq \beta_\kappa$$

For instance, for ACS from Figure 4.1, we have $\beta_{2|s} = \{2\}$ if $s_0 = \{0, 2\}$; whereas we have $\beta_{2|s} = \{2, 10, 1000\}$ if $s_0 = \{0, 10\}$, because we know that 2 is an ancestor of itself under ACS, even when 2 is only observed via 10 and is not selected in s_0 itself.

Theorem 4.1. *Given graph sampling from $G = (U, A)$ defined by* (1.2)*, where each κ for θ defined by* (1.1) *has a positive probability to be observed in the sample graph G_s, the strategy BIGS-IWE defined for $\mathcal{B}' = (F, \Omega; H')$ subjected to* (2.3) *is unbiased for θ, where the ancestor set β'_κ (of κ in $\mathcal{B}'$) satisfies*

$$\emptyset \neq \beta'_\kappa \subseteq \beta_{\kappa|s}$$

Proof. By stipulation, every κ in Ω has a positive probability to be sampled. The term $\sum_{i \in \beta'_\kappa \cap s_0} W_{i\kappa}\, y_\kappa / \pi_i$ in the IWE (2.2) can be specified for each κ separately from the others in Ω. Given the sample ancestors $\beta_{\kappa|s}$ observed on the first occasion when κ is sampled, the non-empty subset β'_κ is known for repeated sampling afterwards. Conditional on the chosen β'_κ and subjected to (2.3), the expectation of $\sum_{i \in \beta'_\kappa \cap s_0} W_{i\kappa}\, y_\kappa / \pi_i$ is y_κ. □

Thus, if the OP of graph sampling from G satisfies the condition (ii) of Theorem 3.1 directly, then one can apply the strategy BIGS-IWE using any fixed non-empty subset of β_κ, and make inference conditionally. Whereas, if the OP of graph sampling from G does not satisfy the condition (ii) of Theorem 3.1, then by Theorem 4.1 one can apply the strategy BIGS-IWE using any fixed non-empty subset of $\beta_{\kappa|s}$ that is observed on the first occasion when κ is sampled, and make inference conditionally.

For ACS from Figure 4.1 above, we need to choose β'_κ based on $\beta_{\kappa|s}$. However, the choice $\beta_2^* = \{2\}$ can actually be determined *a priori*, whereas the choice $\beta_2^\dagger = \{10, 1000\}$

can only be identified if 2 is first time sampled together with its NT network $\{10, 1000\}$. Moreover, in the latter case, where $\beta_{2|s} = \{2, 10, 1000\}$, one can also choose $\beta'_\kappa = \beta_{\kappa|s}$, in which case the estimator $\hat{\theta}_y$ is actually the same as the HTE under $\mathcal{B}$.

Table 4.1 lists the details of the different trategies with respect to BIGS from $\mathcal{B}$, $\mathcal{B}^*$ and $\mathcal{B}^\dagger$. The respective observed sample Ω_s is given in addition to the initial sample s_0. Under the strategy $(\mathcal{B}, \hat{\theta}^*_{HT})$, 2 is given in italic in the 5 samples where it is observed but ineligile for estimation. The probability that it is eligible is 2/5, which is the same as its sample inclusion probability under BIGS from $\mathcal{B}^*$.

Table 4.1: Strategies using BIGS for ACS from G in Figure 4.1.

	$(\mathcal{B}, \hat{\theta}^*_{HT})$ or $(\mathcal{B}, \hat{\theta}_y)$			$(\mathcal{B}^*, \hat{\theta}_y)$		$(\mathcal{B}^\dagger, \hat{\theta}_y)$	
s_0	Ω_s	$\hat{\mu}^*_{HT}$	$\hat{\mu}_y$	Ω_s	$\hat{\mu}_y$	Ω_s	$\hat{\mu}_y$
1,0	1,0	0.500	0.500	1,0	0.500	1,0	0.500
1,2	1,2	1.500	0.944	1,2	1.500	1	0.500
0,2	0,2	1.000	0.444	0,2	1.000	0	0.000
1,10	1,10,*2*,1000	289.071	289.516	1,10,1000	289.071	1,10,2,1000	289.643
1,1000	1,1000,*2*,10	289.071	289.516	1,1000,10	289.071	1,1000,2,10	289.643
0,10	0,10,*2*,1000	288.571	289.016	0,10,1000	288.571	0,10,2,1000	289.143
0,1000	0,1000,*2*,10	288.571	289.016	0,1000,10	288.571	0,1000,2,10	289.143
2,10	2,10,1000	289.571	289.016	2,10,1000	289.571	2,10,1000	289.143
2,1000	2,1000,10	289.571	289.016	2,1000,10	289.571	2,1000,10	289.143
10,1000	10,1000,*2*	288.571	289.016	10,1000	288.571	10,1000,2	289.143
Var.		17418.4	17482.4		17418.4		17533.7

Apart from 2 in italics, the observed sample Ω_s is always the same under both the strategies $(\mathcal{B}, \hat{\theta}^*_{HT})$ and $(\mathcal{B}^*, \hat{\theta}_y)$. The two estimators still differ regarding the RB method. The difference hinges on the last sample $s_0 = \{10, 1000\}$. Under $(\mathcal{B}, \hat{\theta}^*_{HT})$, the same sample (including 2) is also observed from $s_0 = \{2, 10\}$ or $\{2, 1000\}$, but the estimate $\hat{\theta}^*_{HT}$ differs because 2 is unused when $s_0 = \{10, 1000\}$. The RB method yields $\hat{\mu}^*_{HTRB} = 289.238$ given $\Omega_s = \{2, 10, 1000\}$. In contrast, under the strategy $(\mathcal{B}^*, \hat{\theta}_y)$ the estimate of μ is unchanged by the RB method, because the observed sample Ω_s from $s_0 = \{10, 1000\}$ differs to that from $s_0 = \{2, 10\}$ or $\{2, 1000\}$.

The strategy $(\mathcal{B}^\dagger, \hat{\theta}_y)$ can be identified in 70% of the possible ACS sample graphs, under which 2 is ineligible (not included in Ω_s) given $s_0 = \{1, 2\}$ or $\{0, 2\}$. The inclusion probability of 2 is raised to 7/10, the same as 10 or 1000. On the same occasions, one can also adopt $\beta'_2 = \{2, 10, 1000\}$ which coincides with β_2 in this particular case, denoted by $(\mathcal{B}, \hat{\theta}_y)$, where the inclusion probability of 2 is further raised to 9/10. Neither $(\mathcal{B}^\dagger, \hat{\theta}_y)$ nor $(\mathcal{B}, \hat{\theta}_y)$ is a good choice for ACS, because 2 is below the threshold by definition, and the resulting sampling variances are somewhat larger than by $(\mathcal{B}^*, \hat{\theta}_y)$.

4.2 An example of spatial two-stage ACS

The left plot in Figure 4.5 gives the setting of two-stage ACS considered by Thompson (1991, p. 1104). Each vertical strip is a primary sampling unit, and each grid a secondary sampling unit. Given a strip selected at the first stage, all the grids belonging to it are searched for the species. Next, neighbouring grids to those with species are searched, and so on. Thus, ACS is applied at the second stage, which is terminated once no more non-empty grids (with species) are found in this way. An edge grid is an empty grid (without species) which is contiguous to one or more non-empty grids.

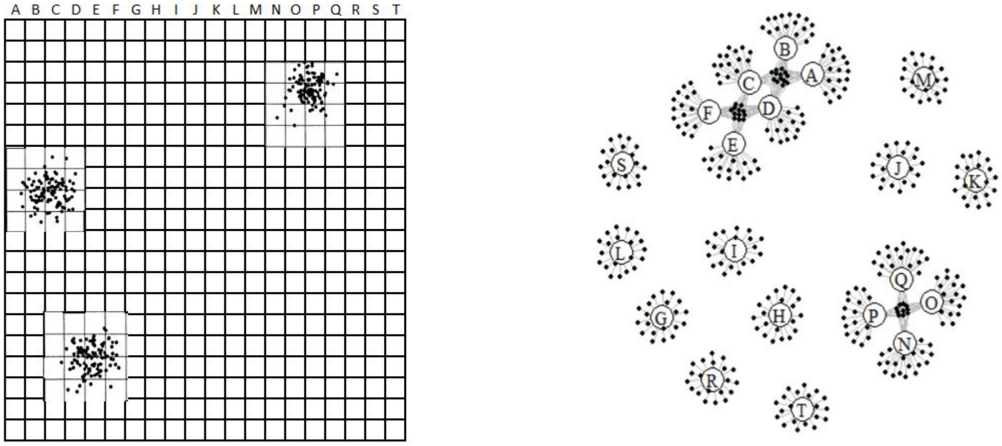

Figure 4.5: Two-stage ACS: left; BIGS representation, right.

The right plot in Figure 4.5 gives the BIG by modifying (3.1), similarly to $\mathcal{B}^*$ in Figure 4.3, where F consists of the strips and Ω the grids. The condition (ii) of Theorem 3.1 is thereby ensured. Each big node marked by a capital letter denotes a strip, the small nodes denote the grids. There are 10 star-like subgraphs, where a strip is adjacent to its 20 empty grids, which are observed under BIGS only if this strip is selected in s_0. The small nodes (of grids, darker in shade) that are adjacent to four big nodes (of strips) form a cluster of non-empty grids, which are all observed if any of the four strips are selected in s_0. There are three such clusters of non-empty grids. Finally, each of the 10 strips that contains non-empty grids is also adjacent to the rest of its empty grids (lighter in shade). Thus, an edge grid is not adjacent to its neighbour strip in $\mathcal{B}^*$, although the former can be observed via the latter under the two-stage ACS design.

Notice that, in this case, one could also use $\mathcal{B}$ by (3.1), where an edge grid is adjacent to its neighbour strip. This is because the y-value is 0 of any empty grid, so it does not contribute to the y-total estimator, whether or not one is able to calculate its inclusion probability. But this is not a generally feasible strategy.

The total of interest in terms of (1.1) is given by

$$\theta = \sum_{\kappa \in \Omega} y_\kappa = \sum_{i \in F} y_i = 326$$

The initial sample s_0 of strips are obtained by SRS, with $m = |s_0|$. Consider

- the HTE $\hat{\theta}$ only based on the m stripes directly selected in s_0;
- the HTE $\hat{\theta}_y$ by (2.6) under BIGS from $\mathcal{B}^*$, where the inclusion probability of a non-empty grid κ is $\pi_{(\kappa)} = 1 - \binom{16}{m} / \binom{20}{m}$, given $|F| = 20$;

- the HH-type multiplicity estimator $\hat{\theta}_{z\beta}$ with the weights $\omega_{i\kappa} = 1/|\beta_\kappa|^{-1}$;
- the HH-type estimator $\hat{\theta}_{z\alpha 1}$ given the PIDA weights (2.9) with $\gamma = 1$, where $|\alpha_i| = 2$ for strips C and D, and $|\alpha_i| = 1$ for the other non-empty strips.

Notice that additional effort is needed for $\hat{\theta}_{z\alpha 1}$, where one must search for possible non-empty grids belonging to each strip encountered during ACS, although one does not need to survey them to obtain the associated y-values.

Table 4.2: Standard errors given $m = |s_0|$

m	$\hat{\theta}$	$\hat{\theta}_y$	$\hat{\theta}_{z\beta}$	$\hat{\theta}_{z\alpha 1}$
1	457	356	356	329
2	315	236	245	226
3	250	179	194	180
4	210	142	163	151
5	182	116	141	131
6	160	95	125	115
7	143	78	111	103
8	128	63	100	92
9	116	51	90	83
10	105	40	82	75

The standard errors are presented in Table 4.2 for $m = 1, ..., 10$. The estimator $\hat{\theta}$ without ACS at the 2nd-stage has the largest standard error given any m. The HTE $\hat{\theta}_y$ is more efficient than both $\hat{\theta}_{z\beta}$ and $\hat{\theta}_{z\alpha 1}$ as m increases; the estimator $\hat{\theta}_{z\alpha 1}$ is the most efficient if $m < 3$, and it is always more efficient than the multiplicity estimator $\hat{\theta}_{z\beta}$. In addition to $\hat{\theta}_{z\beta}$, there are infinite ways of constructing the weights ω_{ik} for $\hat{\theta}_z$ by (2.7), which does not require additional effort like $\hat{\theta}_{z\alpha 1}$. Some of them may well be more efficient than the estimators illustrated here.

4.3 ACS from unknown graphs

In spatial ACS above, the population graph $G = (U, A)$ is known, but the associated values $y_U = \{y_i : i \in U\}$ are unknown. Consider a setting of epidemiological study, where only U is known but not A. Let $N = |U|$. Let $y_i = 1$ if person $i \in U$ is a *case*, and $y_i = 0$ otherwise. Let the population *prevalence* be

$$\mu = \theta/N \qquad \text{where} \qquad \theta = \sum_{i \in U} y_i$$

Provided the virus is transmitted via personal contacts, let $(ij), (ji) \in A$ if relevant contact exists between individuals i and j, such that A contains all the relevant contacts in $U \times U$. The population graph $G = (U, A)$ is undirected and simple.

4.3.1 FPS from U

Stratified multistage sampling from U is a standard FPS design. Denoted by Ω all the ultimate sampling units (USUs), which can be individual, household, building or even neighbourhood. When the USU is a cluster of individuals, it can be selected via the individuals initially selected. Let $y_\kappa = \sum_{i\in\kappa} y_i$ for each $\kappa \in \Omega$, so that $\theta = \sum_{\kappa\in\Omega} y_\kappa$.

BIGS can be used to represent FPS. Let $\mathcal{B} = (U, \Omega; H)$ by (3.1), where $(i\kappa) \in H$ if the selection of individual i leads to that of USU κ. Notice that the edges in H here represent the observational links determined by the choice of the sampling frame and USU, not the contacts A in the population. This is the situation of BIGS with $|\alpha_i| \equiv 1$ considered in Section 2.3. There is little difference of efficiency between the HTE and the HH-type estimator with the PIDA weights (2.9).

4.3.2 ACS from G with unknown A

Epidemic outbreaks can be clustered via the contacts in A but with a low population prevalence μ. Although A is unknown, ACS is worth considering, since intuitively one would like to increase the sample yield of cases with $y_i = 1$ relatively 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 $\pi_i = \Pr(i \in s_0)$. By *adaptive network tracing*, 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 $\pi_{(i)}$ be the inclusion probability in the final sample. Since any case i that is in contact with other cases can be sampled by adaptive network tracing, even when it is not selected in s_0 initially, we achieve $\pi_{(i)} > \pi_i$.

This is a special case of ACS with binary y_i. For each initial sample case $i \in s_0$, with $y_i = 1$, include all the adjacent individuals in G,

$$\nu_i = \{j : (ij) \in A\}$$

Let $s_1 = \bigcup_{i\in s_0} \nu_i \setminus s_0$ be the 1st-wave sample. Repeat the OP for everyone in s_1, which result in the 2nd-wave sample $s_2 = \bigcup_{i\in s_1} \nu_i \setminus (s_0 \cup s_1)$. The procedure is repeated, till it results in an empty wave sample, say, $s_q = \emptyset$. The seed (and final) sample is

$$s = s_0 \cup s_1 \cup \cdots \cup s_{q-1}$$

The sample s can be divided into cases and noncases. The cases can be partitioned into *case networks*, where the individuals in the same case network all have $y_i = 1$ and are connected to each other via the edges in A. For any case i, the ACS sample inclusion probability $\pi_{(i)}$ is the probability that the intersection of s_0 and the network of i is non-empty. Each noncase with $y_i = 0$ may be directly selected in s_0 or via its adjacent case network. Its inclusion probability $\pi_{(i)}$ cannot be calculated generally. However, this does not matter here, since a noncase contributes 0 to the estimation of θ regardless.

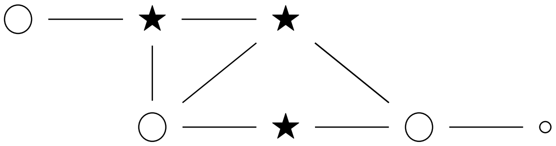

Figure 4.6: Illustration of case (★) edge node (◯) and other noncase (∘) in G

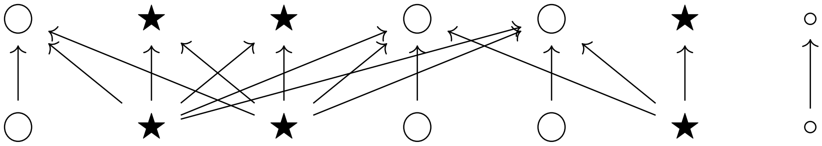

Figure 4.7: BIG by (3.1) corresponding to Figure 4.6.

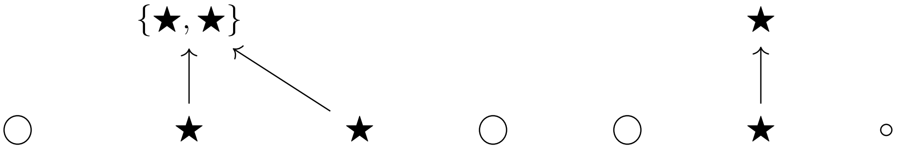

Figure 4.8: BIG with case networks as study units for ACS from Figure 4.6.

Figure 4.6 illustrates two networks of cases (★), their edge nodes (◯) and another noncase (∘) in G. Figure 4.7 shows the corresponding $\mathcal{B} = (F, \Omega; H)$ by (3.1), where $F = \Omega = U$ and H contains all the observational links under ACS from G. In particular, any ★ would lead to its adjacent ◯ but not vice versa. Finally, Figure 4.8 shows the BIG, where $F = U$ and Ω consists only of the case networks, after removing the non-contributing noncases from the consideration.

In terms of the BIG as illustrated in Figure 4.8, let κ be the network of case i, with nodes β_κ and $n_\kappa = |\beta_\kappa|$. The HTE of θ can be given by

$$\hat{\theta}_y = \sum_{i \in s} \frac{y_i}{\pi_{(i)}} = \sum_{\kappa \in \Omega} \frac{n_\kappa}{\pi_{(\kappa)}} \tag{4.1}$$

where $\pi_{(\kappa)} = \pi_{(i)}$ for any $i \in \beta_\kappa$. We have

$$V(\hat{\theta}_y) = \sum_{\substack{i \in U \\ y_i = 1}} \sum_{\substack{j \in U \\ y_j = 1}} \Big(\frac{\pi_{(ij)}}{\pi_{(i)}\pi_{(j)}} - 1 \Big) = \sum_{\substack{\kappa \in \Omega \\ n_\kappa > 0}} \sum_{\substack{\ell \in \Omega \\ n_\ell > 0}} \Big(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)}\pi_{(\ell)}} - 1 \Big) n_\kappa n_\ell \tag{4.2}$$

since only $y_i y_j = 1$ contribute to the summation of (ij) over $U \times U$.

For the HH-type estimator, we have $z_i = y_i$ for any $i \in \beta_\kappa$ since $|\alpha_i| = n_\kappa$ for each case i in network κ. Provided $\pi_{(\kappa)} \approx \sum_{i \in \beta_\kappa} \pi_i$, there is little difference to the HTE

The network-exhaustive OP of ACS could be an issue if a network is too large to be surveyed completely due to practical reasons. If it is possible to measure ξ_{ij} as the *strength* of $(ij) \in A$, then one may define *adaptively* the neighbourhood of i to be

$$\nu_i^* = \{ j : (ij) \in A,\, \xi_{ij} > \xi_0 \}$$

for a chosen threshold ξ_0, and include ν_i^* if $y_i = 1$. To distinguish, the corresponding sampling method may be referred to as *doubly adaptive cluster sampling*, since it is based on two threshold values $y_i > 0$ and $\xi_{ij} > \xi_0$.

Imposing a maximum number of waves, say q, is another way to curtail large networks. The sampling is terminated after the q-th wave, even if $s_q \neq \emptyset$, yielding the sample $s = \sum_{r=0}^{q} s_r$. This is a *q-wave adaptive snowball sampling* design. Snowball sampling will be considered in Chapter 5.

4.3.3 Some simulation results

Stratified multistage sampling from U is standard FPS, but it may be less efficient than ACS or its modifications. It is possible to consider a combined approach below.

- Deploy a stratified multistage sampling design to secure a baseline precision for the estimation of θ and μ. Denote by $\tilde{s}$ the selected sample.
- Apply ACS, or its modified version, to a subsample $s_0 \subset \tilde{s}$, to collect additional data about cases and contacts. For instance, s_0 may consist of the asymptomatic cases.
- For estimation of θ and μ one may combine the two samples. The data collected under ACS are useful for epidemiological modelling and analysis in any case.

Moreover, to inform the decision, one may use simulations to study the salient aspects of the sampling design, as illustrated below.

Size-biased sampling and adaptive network tracing

Let η be the odds of case selection in the initial sample s_0, which is defined as the ratio of the probability that a case is included in s_0 against that of a noncase. Initial sampling is *size-biased* if $\eta \neq 1$, positively so if $\eta > 1$. Moreover, adaptive network tracing is applied to the cases under ACS, which yields a sample of case networks. Positive size-biased initial sampling and adaptive network tracing are potentially two key points to an efficient sampling design, in light of two simple observations below.

- The most efficient initial sample of size $m \geq N\mu$ is the one which includes all the cases with probability 1, where $\eta = \infty$ and the sampling variance is 0.
- A case network is nearly always included by adaptive network tracing, if the probability is virtually 1 that it intersects the initial sample s_0, even when the probability is far below 1 for all these cases to be included in s_0, as long as $m \ll N$ and $\eta \ll \infty$.

Table 4.3 presents some results for the RE of ACS based on adaptive network tracing, defined as the ratio of the variance of the HTE under ACS against that based on the initial sample s_0, which is either selected by SRS ($\eta = 1$) or sized-biased sampling with $\eta = 2$. All the cases in the population are divided into networks, which have the same size c. We observe the following in particular.

Table 4.3: ACS given equal-size (c) case networks in population of size $N = 10^5$, $\theta = 10^3$ and prevalence $\mu = 0.01$. Initial sample of size m by SRS ($\eta = 1$) or size-biased sampling ($\eta = 2$), ACS with sample size $n = |s|$ by adaptive network tracing.

SRS ($\eta = 1$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.31	1631	0.24	0.58	1085	0.31	0.96	1010	0.31	0.99
1630	0.24	2423	0.15	0.40	1766	0.24	0.93	1646	0.24	0.99
2420	0.20	3306	0.10	0.23	2614	0.19	0.89	2443	0.20	0.99
5000	0.14	5944	0.02	0.03	5352	0.12	0.79	5048	0.14	0.97
10000	0.09	10900	0.00	0.00	10551	0.07	0.59	10090	0.09	0.95

Size-biased ($\eta = 2$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.22	1840	0.13	0.32	1160	0.21	0.91	1020	0.22	0.99
5000	0.09	5901	0.00	0.00	5549	0.07	0.60	5090	0.09	0.95
10000	0.06	10802	0.00	0.00	10692	0.04	0.31	10159	0.06	0.89

- ACS is increasingly more efficient than SRS as m increases, if one compares the CVs of SRS with m and ACS with $E(n)$, where $m \approx E(n)$. The gain is more pronounced given large networks. Given initial SRS of size 1×10^4, ACS requires about 900 extra individuals, by which the sampling variance is reduced to 0.00, because any case network intersects s_0 almost certainly. The reduction is quicker given initial size-biased sampling, where the variance is already 0.00 at $m = 0.5 \times 10^4$.

- ACS has basically no gains given only small case networks with $k = 2$, where size-biased sampling would be the chief means for reducing variance. Given $\eta = 2$ the variance of the initial sample estimator is about halved given any m in Table 4.3.

In summary, size-biased sampling and adaptive network tracing can enhance each other, generating extra gains when they are applied jointly.

The inclusion probabilities $\pi_{(\kappa)}$ and $\pi_{(\kappa\ell)}$ are easy to compute under initial SRS. For unequal probability sampling of s_0, the exclusion probabilities $\bar{\pi}_{\beta_\kappa}$ are usually unknown if $|\beta_\kappa| > 1$. When the sampling fraction is low, it is convenient to treat the initial sampling as if it were Poisson sampling, where the individuals are independently selected. It has been verified empirically that the approximation holds well in the simulation settings here, including the highest sampling fraction 10%.

Population of households

Households can be envisaged as social bubbles, within which contact is hard to avoid. Denote by $\mathbb{G} = (\mathbb{H}, \mathbb{A})$ the population graph, where $\mathbb{H}$ consists of all the households, and $\mathbb{A}$ the contacts between any two households via their members. Size-biased sampling and adaptive network tracing in $\mathbb{G}$ follow the same definition as in $G = (U, A)$, but the actual design effects will differ between the two set-ups.

Table 4.4 presents some results for the RE of ACS based on adaptive network tracing,

Table 4.4: ACS given equal-size (c) case networks in household population of size $N = 10^5$, $\theta = 10^3$ and prevalence $\mu = 0.01$. Initial sample of size m by SRS ($\eta = 1$) or size-biased sampling ($\eta = 2$), ACS with sample size $n = |s|$ by adaptive network tracing.

SRS ($\eta = 1$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.35	1628	0.24	0.45	1086	0.31	0.77	1010	0.34	0.89
5000	0.16	5944	0.02	0.03	5351	0.13	0.63	5048	0.14	0.87
10000	0.11	10900	0.00	0.00	10552	0.07	0.49	10090	0.10	0.84

Size-biased ($\eta = 2$)		ACS, $c = 100$			ACS, $c = 10$			ACS, $c = 2$		
m	CV	$E(n)$	CV	RE	$E(n)$	CV	RE	$E(n)$	CV	RE
1000	0.25	1844	0.13	0.26	1162	0.22	0.71	1019	0.23	0.86
5000	0.11	5901	0.00	0.00	5547	0.08	0.47	5089	0.10	0.85
10000	0.07	10802	0.00	0.00	10691	0.04	0.25	10159	0.06	0.80

which are comparable to those of Table 4.3. The only difference is that sampling and network tracing are from a population of households instead of persons. The stipulated household size distribution is $(0.38, 0.30, 0.12, 0.20)$ for household size $(1, 2, 3, 4)$, among both the households of cases and the households of noncases. The differences of $E(n)$ in Table 4.3 and 4.4 reflect the magnitude of simulation error in these results.

- The variances are larger than the corresponding ones under sampling of individuals. However, the increases under ACS are smaller, such that the relative efficiency gains by adaptive network tracing are actually increased compared to sampling of persons, where the RE is appreciable even when the networks are small, e.g. $k = 2$.

- The RE of size-biased initial sampling is similar to that in Table 4.3.

4.4 ACS from unknown dynamic graphs

The population graph in epidemiological studies is dynamic, denoted by $G_t = (U_t, A_t)$, for time point $t = 1, 2, ...$ Even when U_t is fixed, the contacts A_t that are relevant for the time point t will be dynamic, unless there is a strict lockdown.

4.4.1 Some basic designs

There are more than one possible definition of U_t, three of which are as below.

a. One may choose to keep $U_t = U$ fixed, as long as one has the knowledge of U.

b. Let U_F be a known population associated with a fixed sampling frame F. One may let U_t be the union of U_F and those that can be linked to U_F via A_t.

c. One can define the population recursively, where U_t is the union of U_{t-1} and those that can be linked to it via A_t, and U_1 is either U by (a) or U_F by (b).

Panel The sample once selected is fixed over time. This design accommodates only the target population definition (a) above. When the change $\mu_2 - \mu_1$ is estimated based on two independent samples at $t = 1, 2$, the variance is the sum of those of $\hat{\mu}_1$ and $\hat{\mu}_2$. The variance can be reduced, provided the panel design induces a positive correlation between $\hat{\mu}_1$ and $\hat{\mu}_2$. Now that prevalence is a proportion, a more direct comparison can be given as follows. Suppose that based on two independent samples of the same size n, the sampling variance of the change estimator is

$$V(\hat{\mu}_2 - \hat{\mu}_1) = \frac{1}{n}\big(\mu_2(1-\mu_2) + \mu_1(1-\mu_1)\big) \doteq \frac{1}{n}\big(\mu_2 + \mu_1\big)$$

as long as $1 - \mu_2 \doteq 1$ and $1 - \mu_1 \doteq 1$. Suppose $N_2 = N_1$ and $\mu_2 = \mu_1$ for simplicity. Let λ_+ be the proportion of the new cases over time, and λ_- that of the closed cases, where $\lambda_+ \leq \mu_2$ and $\lambda_- \leq \mu_1$ by definition. Let $e_i = 0$ if individual i has no change of case status, $e_i = 1$ if i becomes a case, and $e_i = -1$ if i becomes a closed case, such that $\mu_2 - \mu_1$ is the population mean of e_i. Based on a panel of size n, it can be shown that

$$V_{panel}(\hat{\mu}_2 - \hat{\mu}_1) \doteq \frac{1}{n}(\lambda_+ + \lambda_-) \leq \frac{1}{n}\big(\mu_2 + \mu_1\big) = V(\hat{\mu}_2 - \hat{\mu}_1)$$

Panel ACS Only the initial sample s_0 is fixed over time, but the sample $s(t)$ obtained based on s_0 and A_t via adaptive network tracing varies with t. This design is applicable to either the target population definition (a) or (b) above.

Iterative ACS Let $s(t)$ be the sample by ACS based on s_0 and A_t for time t, with inclusion probabilities $\pi_i(t)$ and $\pi_{ij}(t)$ for $i, j \in s(t) \subset U_t$, and let $s(t+1)$ be given by ACS based on A_{t+1} for time $t+1$ and starting from

$$s_{0t} = s_0 \cup \{\, i \in s(t) \setminus s_0 : y_{i,t} = 1\}$$

Design-based inference of $\mu_{t+1} - \mu_t$ is possible using $\pi_i(t)$ and $\pi_{ij}(t)$, even though one cannot calculate the probabilities $\pi_i(t+1)$ and $\pi_{ij}(t+1)$ for $i, j \in s(t+1) \subset U_{t+1}$. This can accommodate all the target population definitions (a) - (c) above.

4.4.2 Estimation of change

The HTE of change $\nabla_{t,t+1} = \mu_{t+1} - \mu_t$ under the panel design is given by

$$\hat{\nabla}^{panel}_{t,t+1} = \frac{1}{N}\sum_{i \in s_0} \frac{1}{\pi_i}(y_{i,t+1} - y_{i,t})$$

where π_i is the inclusion probability of $i \in s_0$, and s_0 is the panel.

The HTE of $\nabla_{t,t+1}$ under the panel ACS design is given by

$$\hat{\nabla}^{pACS}_{t,t+1} = \frac{1}{N_{t+1}}\sum_{i \in s(t+1)} \frac{y_{i,t+1}}{\pi_i(t+1)} - \frac{1}{N_t}\sum_{i \in s(t)} \frac{y_{i,t}}{\pi_i(t)}$$

where $s(t)$ is the sample at time t by ACS based on s_0 and A_t, and $\pi_i(t)$ is the inclusion probability of $i \in s(t)$, and similarly for $s(t+1)$ and $\pi_i(t+1)$. That is, one applies (4.1) at each time point and take the difference between them. The variance of each HTE follows from (4.2). Similarly for the covariance between them, where we have

$$
\begin{aligned}
Cov(\hat{\mu}_t, \hat{\mu}_{t+1}) &= \frac{1}{N_t N_{t+1}} \sum_{\substack{i \in U_t \\ y_{i,t}=1}} \sum_{\substack{j \in U_{t+1} \\ y_{j,t+1}=1}} \big(\frac{\pi_{(ij)}}{\pi_{(i)}\pi_{(j)}} - 1 \big) y_{i,t} y_{j,t+1} \\
&= \frac{1}{N_t N_{t+1}} \sum_{\substack{\kappa \in \Omega_t \\ n_{\kappa,t}>0}} \sum_{\substack{\ell \in \Omega_{t+1} \\ n_{\ell,t+1}>0}} \big(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)}\pi_{(\ell)}} - 1 \big) n_{\kappa,t} n_{\ell,t+1}
\end{aligned}
$$

The first-order inclusion probabilities can be calculated as usual under either BIGS. For the second-order inclusion probabilities, notice that the two networks κ and ℓ refer to two different time points here, such that one needs to take into account two BIGs, similarly as detailed below for the iterated ACS design.

Under the iterated ACS design, an unbiased estimator of $\nabla_{t,t+1}$ is given by

$$
\hat{\nabla}^{iACS}_{t,t+1} = \frac{1}{N_{t+1}} \Big(\sum_{\substack{i \in s_{0t} \\ y_{i,t}=1}} \frac{y_{i,t+1}}{\pi_i(t)} + \sum_{\substack{i \in s_{0t} \\ y_{i,t}=0}} \frac{y_{i,t+1}}{\pi_i} \Big) - \frac{1}{N_t} \sum_{i \in s(t)} \frac{y_{i,t}}{\pi_i(t)} \tag{4.3}
$$

The two terms in the parentheses form an HH-type estimator of θ_{t+1}, where the values $\{y_{j,t+1} : j \in s(t+1)\}$ are transformed to $\{z_i : i \in s_{0t}\}$. Now that $y = 0$ or 1, we have $z_i = 1$ using the multiplicity weights if $y_{i,t+1} = 1$, and 0 otherwise, so that $z_i = y_{i,t+1}$. Meanwhile, the inclusion probability in s_{0t} differs depending on whether an individual is case or not at t, corresponding to the two terms given above for emphasis, respectively, since s_{0t} is obtained by ACS based on s_0 and A_t.

The estimator (4.3) can also be considered as the HTE based on s_{0t}, with value $y_{i,t+1}/N_{t+1} - y_{i,t}/N_t$ for each $i \in s_{0t}$. The variance follows. The inclusion probability of $i \in s(t)$ has been explained before. Let κ be the network of individual i, where $\beta_\kappa = \{i\}$ if $y_{i,t} = 0$. Let ℓ be that of j. The joint inclusion probability of $i \neq j \in s(t)$ is given by

$$
\pi_{(ij)} = \begin{cases}
\pi_{ij} & \text{if } y_{i,t} = 0,\ y_{j,t} = 0 \\
\pi_i + \pi_{(j)} + \bar{\pi}_{\{i\} \cup \beta_\ell} - 1 & \text{if } y_{i,t} = 0,\ y_{j,t} = 1 \\
\pi_{(i)} + \pi_j + \bar{\pi}_{\beta_\kappa \cup \{j\}} - 1 & \text{if } y_{i,t} = 1,\ y_{j,t} = 0 \\
\pi_{(i)} + \pi_{(j)} + \bar{\pi}_{\beta_\kappa \cup \beta_\ell} - 1 & \text{if } y_{i,t} = 1,\ y_{j,t} = 1
\end{cases}
$$

The estimator $\hat{\nabla}^{pACS}_{t,t+1}$ by panel ACS can be more efficient than $\hat{\nabla}^{panel}_{t,t+1}$ under the panel design, because s_0 is a subsample of either $s(t)$ or $s(t+1)$, and ACS increases the sample inclusion probability of a case. Likewise between $\hat{\nabla}^{iACS}_{t,t+1}$ and $\hat{\nabla}^{panel}_{t,t+1}$. The RE between panel and iterated ACS is undetermined in general. On the one hand, the sample $s(t+1)$ based on s_0 and A_{t+1} is a subsample of that based on s_{0t} and A_{t+1} because $s_0 \subseteq s_{0t} \subseteq s(t)$; on the other hand, the HH-type estimator of $\hat{\mu}_{t+1}$ under iterated ACS may be less efficient than the HTE under panel ACS. The strengths of the contrasting forces depend on how the case networks in A_t and A_{t+1} relate to each other.

4.4.3 Simulation results over two time points

New case networks may emerge from one time point to the next, whilst the existing ones may increase or decrease in their sizes. The speed may be quick or slow, at which a new case network emerges or an existing one grows or shrinks. Some settings over two time points are given in Table 4.5, where both the population size and prevalence are constant, such that the target parameter is $\nabla_{1,2} = 0$ in all the settings. Notice that the networks are all of size 2 at $t = 1$ in the last three settings S1-S3: for those that are not growing, their sizes at $t = 2$ are randomly assigned, subjected to the case total $\theta_2 = 10^3$, such that some of them may simply disappear by chance.

Table 4.5: Populations of constant size $N = 10^5$ and total $\theta = 10^3$ at $t = 1, 2$. With (number, size) of case networks: at $t = 1$, $(\bar{\theta}, c)$ networks; at $t = 2$, $(\bar{\theta}_+, c_+)$ or $(\bar{\theta}_-, c_-)$ existing networks of increasing or decreasing sizes, and $(\bar{\theta}', c')$ emerging networks.

		$t = 1$	$t = 2$		
Setting	Characterisation	$(\bar{\theta}, c)$	$(\bar{\theta}_+, c_+)$	$(\bar{\theta}_-, c_-)$	$(\bar{\theta}', c')$
L1	Large, Quickly Evolving	(10, 100)	(2, 180)	(8, 80)	(0, 0)
L2	Large, Quickly Emerging	(10, 100)	(0, 0)	(10, 80)	(2, 100)
L3	Large, Slowly Emerging	(10, 100)	(0, 0)	(10, 90)	(5, 20)
M1	Medium, Quickly Evolving	(100, 10)	(10, 46)	(90, 6)	(0, 0)
M2	Medium, Quickly Emerging	(100, 10)	(0, 0)	(100, 6)	(10, 40)
M3	Medium, Slowly Emerging	(100, 10)	(0, 0)	(100, 9)	(10, 10)
S1	Small, Quickly Evolving	(500, 2)	(10, 42)	($\leq$490, $\leq$2)	(0, 0)
S2	Small, Quickly Emerging	(500, 2)	(0, 0)	($\leq$500, $\leq$2)	(10, 40)
S3	Small, Slowly Emerging	(500, 2)	(0, 0)	($\leq$500, $\leq$2)	(50, 2)

Table 4.6 presents some simulation results for the settings in Table 4.5. The RE of an adaptive design is calculated against the panel design.

- Overall, from top-right towards bottom-left in Table 4.6, the RE of panel ACS is seen to improve quickly with the three initial values of sample size m, odds of case selection η and case network size c. The RE of iterated ACS improves with m except for S1-S3, but not with η, although it remains more efficient than the panel design as η increases. The improvements are larger for panel ACS than iterated ACS.

- For any given initial network size c, moving between the three patterns of case networks over time, the RE of panel ACS improves less for slowly than quickly changing networks, as m increases. As the initial odds of case selection η increases, the RE of the panel ACS become more uniform across all the patterns.

- Given any combination of initial values of (c, m, η), the RE of iterated ACS becomes more uniform across the three patterns of case networks, as m and η increase.

- Between the two ACS designs, the panel ACS is more efficient given sufficiently large initial sample size m and high odds of case selection η, whereas the iterated ACS is

Table 4.6: Panel, panel ACS and iterated ACS designs for settings in Table 4.5.

	Initial SRS of Size $m = 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}^{panel}_{t,t+1}$)	0.20	0.20	0.14	0.28	0.28	0.14	0.28	0.28	0.13
RE($\hat{\nabla}^{pACS}_{t,t+1}$)	0.71	0.73	0.90	0.89	0.89	0.98	0.89	0.91	0.98
RE($\hat{\nabla}^{iACS}_{t,t+1}$)	0.57	0.60	0.52	0.69	0.70	0.52	0.84	0.85	0.75
	Initial SRS of Size $m = 5 \times 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}^{panel}_{t,t+1}$)	0.09	0.09	0.06	0.12	0.12	0.06	0.12	0.12	0.06
RE($\hat{\nabla}^{pACS}_{t,t+1}$)	0.09	0.11	0.38	0.62	0.63	0.87	0.65	0.64	0.91
RE($\hat{\nabla}^{iACS}_{t,t+1}$)	0.49	0.51	0.49	0.67	0.67	0.53	0.85	0.84	0.76
	Initial Size-biased Sampling ($\eta = 2$) of Size $m = 10^3$								
(SE in 10^{-2})	L1	L2	L3	M1	M2	M3	S1	S2	S3
SE($\hat{\nabla}^{panel}_{t,t+1}$)	0.17	0.17	0.12	0.24	0.24	0.12	0.24	0.24	0.12
RE($\hat{\nabla}^{pACS}_{t,t+1}$)	0.31	0.33	0.41	0.51	0.51	0.50	0.52	0.53	0.51
RE($\hat{\nabla}^{iACS}_{t,t+1}$)	0.70	0.69	0.68	0.79	0.81	0.68	0.89	0.90	0.84

more efficient given small m and initial SRS, especially given slowly changing networks, where the panel ACS does not yield much gain over the panel design.

The improvements of iterated ACS is useful given relatively small m, if positively size-biased sampling is difficult to achieve, e.g. due to a lack of understanding of the relevant risk factors, or a lack of frame data that can be used to effectively vary the initial sample inclusion probability even though the relevant factors are known. Together, the panel and iterated ACS designs seem to complement each other in different settings, offering helpful choices across a wider range of situations than each on its own.

Bibliographic notes

As a feasible strategy for ACS, Thompson (1990) proposes a modified HT-estimator, where an edge grid is used for estimation, only if it is selected in s_0 directly, the probability of which is known from the initial sampling design. This is presented in Section 4.1.2 using $\mathcal{B}$. Thompson (1990) proposes also a modified HH-type estimator based on the multiplicity weights, where an edge grid is used for estimation only if it is selected in s_0 directly. This modified HH-type estimator under BIGS from $\mathcal{B}$ is the same as the unmodified HH-type estimator under BIGS from $\mathcal{B}^*$ in Section 4.1.2.

Zhang and Oguz-Alper (2020) extend the innovative approach of Thompson (1990) and apply IWE using suitably constructed BIG, based on the recognition that one can either modify the sampling method or the estimator in a given sampling strategy. Using $\mathcal{B}$ that includes all the observational links under ACS, one can modify the HT or HH-type estimators, as originally proposed by Thompson (1990). Using the graph $\mathcal{B}^*$ that modifies the observational links under ACS to ensure the ancestry knowledge, one can use the standard HT or HH-type estimator.

The numerical example of two-stage ACS is taken from Zhang and Oguz-Alper (2020). The estimators other than $\hat{\theta}_{z\alpha 1}$ are also considered by Thompson (1991).

The designs for cross-sectional and change estimation in Sections 4.3 and 4.4 have been proposed by Zhang (2020).

Chapter 5

Snowball sampling

Applying incident OP enlarges an initial sample of nodes by their out-of-sample adjacent nodes, the repetition of which resembles the rolling of a snowball. T-wave *Snowball sampling (TSBS)* of a given number of waves is not necessarily network exhaustive. It can be regarded as a probabilistic breadth-first search algorithm in graphs.

5.1 T-wave snowball sampling

TSBS from $G = (U, A)$ is given by applying the T-wave incident OP to an initial sample s_0, as defined in Section 1.3.3. The sample graph $G_s = (U_s, A_s)$ follows from (1.2). For digraphs, the reference set and the sample edges generated by TSBS are given by

$$s_{\text{ref}} = s \times U \qquad \text{and} \qquad A_s = \bigcup_{i \in s} \bigcup_{j \in \alpha_i} A_{ij}$$

where s is the seed sample. For undirected graphs, we have

$$s_{\text{ref}} = s \times U \cup U \times s \qquad \text{and} \qquad A_s = \bigcup_{i \in s} \bigcup_{j \in \alpha_i} (A_{ij} \cup A_{ji})$$

5.1.1 Inclusion probabilities of nodes and edges in G_s

For simplicity, assume $F = U$. By way of introduction consider the inclusion probabilities $\pi_{(i)}$ and $\pi_{(i)(j)}$ of nodes in U_s, and $\pi_{(ij)}$ and $\pi_{(ij)(hl)}$ of edges in A_s.

The wave samples $s_0, ..., s_T$ are disjoint under TSBS. For a given sample edge $(ij) \in A_s$ from digraphs, there can only be one particular wave t, where $t \leq T - 1$ and $i \in s_t$. For any $i \in U$, let $\beta_i^{[0]} = \{i\}$, and let its *t-th wave ancestors* be

$$\beta_i^{[t]} = \beta(\beta_i^{[t-1]}) \setminus \bigcup_{r=0}^{t-1} \beta_i^{[r]} \quad \text{for} \quad t > 0$$

which consists of the nodes that would lead to i in t waves but not sooner. Notice that

$\beta_i^{[0]}, \beta_i^{[1]}, ..., \beta_i^{[T]}$ are disjoint. We have

$$\pi_{(i)} = 1 - \bar{\pi}_{B_i} \qquad \text{for} \quad B_i = \bigcup_{t=0}^{T} \beta_i^{[t]}$$

$$\pi_{(ij)} = 1 - \bar{\pi}_{B_{ij}} \qquad \text{for} \quad B_{ij} = \bigcup_{t=0}^{T-1} \beta_i^{[t]}$$

The respective 2nd-order inclusion probabilities follow as $\pi_{(i)(j)} = 1 - \bar{\pi}_{B_i} - \bar{\pi}_{B_j} + \bar{\pi}_{B_i \cup B_j}$ and $\pi_{(ij)(hl)} = 1 - \bar{\pi}_{B_{ij}} - \bar{\pi}_{B_{hl}} + \bar{\pi}_{B_{ij} \cup B_{hl}}$.

For a given sample edge $(ij) \in A_s$ from undirected graphs, there can only be one particular wave t, where $t \leq T-1$ and *either* $i \in s_t$ *or* $j \in s_t$. For any $i \in U$, let

$$\nu_i = \{j \in U : a_{ij} + a_{ji} > 0\}$$

be the set of its adjacent nodes. Let $\nu_i^{[0]} = \{i\}$, and let its *t-th wave ancestors* be

$$\nu_i^{[t]} = \nu(\nu_i^{[t-1]}) \setminus \bigcup_{r=0}^{t-1} \nu_i^{[r]} \quad \text{for} \quad t > 0$$

which are the nodes that would lead to i in t waves but not sooner. Notice that $\nu_i^{[0]}, \nu_i^{[1]}, ..., \nu_i^{[T]}$ are disjoint. We have

$$\pi_{(i)} = 1 - \bar{\pi}_{R_i} \qquad \text{for} \quad R_i = \bigcup_{t=0}^{T} \nu_i^{[t]} \tag{5.1}$$

$$\pi_{(ij)} = 1 - \bar{\pi}_{R_{ij}} \qquad \text{for} \quad R_{ij} = \bigcup_{t=0}^{T-1} (\nu_i^{[t]} \cup \nu_j^{[t]}) \tag{5.2}$$

The respective 2nd-order inclusion probabilities follow as $\pi_{(i)(j)} = 1 - \bar{\pi}_{R_i} - \bar{\pi}_{R_j} + \bar{\pi}_{R_i \cup R_j}$ and $\pi_{(ij)(hl)} = 1 - \bar{\pi}_{R_{ij}} - \bar{\pi}_{R_{hl}} + \bar{\pi}_{R_{ij} \cup R_{hl}}$.

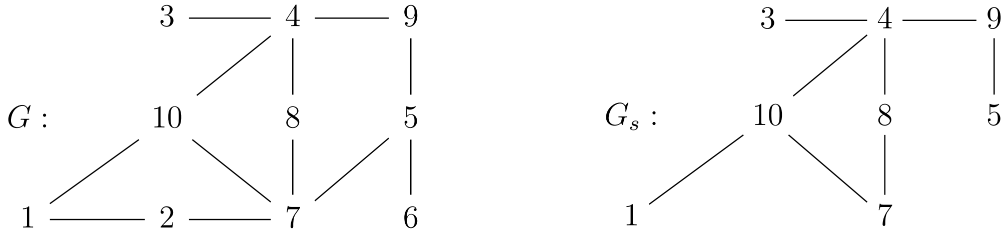

Figure 5.1: 2SBS yielding G_s from G given $s_0 = \{3, 4\}$.

Figure 5.1 shows the sample graph $G_s = (U_s, A_s)$ by 2SBS given $s_0 = \{3, 4\}$, from the population graph $G = (U, A)$. The 1st and 2nd wave samples are $s_1 = \{8, 9, 10\}$ and $s_2 = \{1, 5, 7\}$. The seed sample is $s = \{3, 4, 8, 9, 10\}$. The node inclusion probabilities $\pi_{(i)}$ are given by (5.1), where there are 5 distinct values, and the edge inclusion probabilities $\pi_{(ij)}$ by (5.2), where there are 4 distinct values. These can be verified by enumeration over all the 45 possible initial samples given $|s_0| = 2$. For example, the node 3 has

$\nu_3^{[0]} = \{3\}$, $\nu_3^{[1]} = \{4\}$ and $\nu_3^{[2]} = \{8, 9, 10\}$, such that $R_3 = \{3, 4, 8, 9, 10\}$. Under SRS of s_0 with $|s_0| = 2$, the exclusion probability of R_3 from s_0 is $\binom{5}{2}/\binom{10}{2} = 2/9$, such that $\pi_{(3)} = 1 - 2/9 = 7/9$, whereas $\pi_3 = \Pr(3 \in s_0) = 1/5$.

5.1.2 Arbitrary M with $|M| \geq 2$ given $s_{\mathbf{ref}} = s \times U \cup U \times s$

Let $M = \{i_1, i_2, ..., i_q\} \subset U$ with $|M| = q$. To identify the motif $\kappa = [M]$, there can be at most one node in M that belongs to the last wave sample s_T. Let $M^{(h)} = M \setminus \{i_h\}$ be the subset obtained by dropping i_h from M, for $h = 1, ..., q$. We have

$$\begin{aligned}\pi_{(\kappa)} &= \Pr\big(M^{(1)} \subseteq s \text{ or } M^{(2)} \subseteq s \text{ or } \cdots \text{ or } M^{(k)} \subseteq s \text{ or } M \subseteq s\big) \\ &= \sum_{h=1}^{q} \Pr\big(M^{(h)} \subseteq s) - (q-1)\Pr(M \subseteq s\big) \qquad (5.3)\end{aligned}$$

This follows from applying the inclusion-exclusion calculus of the probability of union of events. In the case here, we have $\big(M^{(h)} \subseteq s\big) \cap \big(M \subseteq s\big) = \big(M \subseteq s\big)$ for any h, and $\big(M^{(h)} \subseteq s\big) \cap \big(M^{(l)} \subseteq s\big) = \big(M \subseteq s\big)$ for $h \neq l$, such that all the terms apart from $\sum_{h=1}^{q} \Pr\big(M^{(h)} \subseteq s\big) + \Pr\big(M \subseteq s\big)$ are proportional to $\Pr\big(M \subseteq s\big)$. The sum of these other terms, $\big(1-(q+1)\big)\Pr\big(M \subseteq s\big)$, can be obtained from matching them to the corresponding terms in the identity $(1-1)^{q+1} = 0$, which yields (5.3).

Next, we have

$$\begin{aligned}&\Pr\big(M \subseteq s\big) = \sum_{L \subseteq M} (-1)^{|L|} \bar{\pi}(L) \\ &\bar{\pi}(L) = \Pr(L \cap s = \emptyset) = \Pr(R_L \cap s_0 = \emptyset) = \bar{\pi}_{R_L} = \sum_{D \subseteq R_L} (-1)^{|D|} \pi_D\end{aligned}$$

where $\bar{\pi}(L)$ is the seed-sample exclusion probability of L, including $\bar{\pi}(\emptyset) = 1$ for $L = \emptyset$, and $R_L = \bigcup_{i \in L} R_i$ with R_i in (5.1), and π_D is the joint inclusion probability of the nodes D in the initial sample s_0. Similarly for $\Pr\big(M^{(h)} \subseteq s\big)$, where $h = 1, ..., q$.

For $M \subset U$ and $M' \subset U$, joint observation of $\kappa = [M]$ and $\kappa' = [M']$ requires at most one node i in s_T if $i \in M \cap M'$, or at most two nodes i, j in s_T if $i \in M \setminus M'$ and $j \in M' \setminus M$. Let $\tilde{M} = M \cup M'$. Define subset $\tilde{M}^{(i)}$ for each $i \in M \cap M'$, and $\tilde{M}^{(ij)}$ for each pair of $i \in M \setminus M'$ and $j \in M' \setminus M$. The joint inclusion probability $\pi_{(\kappa\kappa')}$ is the probability that at least one of these subsets of $\tilde{M}$ is in the seed sample s, which can be obtained by the inclusion-exclusion calculus similarly as (5.3).

5.1.3 Arbitrary M with $|M| \geq 2$ using $s^*_{\mathbf{ref}} = s \times s$

By dropping the last wave sample s_T, we ensure that the motif $[M]$ is observed if $M \subset s$, because incident OP under TSBS implies induced OP in s. That is, let $G_s = (U_s, A_s)$ be the sample graph of TSBS, with $s_{\text{ref}} = s \times U \cup U \times s$. Let $G^*_s = (U^*_s, A^*_s)$ be the reduced sample graph obtained from dropping s_T, given by $s^*_{\text{ref}} = s \times s$, where

$$A^*_s = A_s \setminus \{(ij) : i \in s, j \in s_T\} \quad \text{and} \quad U^*_s = U_s \setminus s_T = s$$

Thus, A^*_s contains all the edges between any $i, j \in s$ in the population graph G, and G^*_s is the same sample graph that is obtained from s by induced observation directly. It follows that $\pi^*_{(\kappa)}$ with respect to s^*_{ref} is directly given by

$$\pi^*_{(\kappa)} = \Pr\big(M \subseteq s\big) \tag{5.4}$$

For $M \subset s$ and $M' \subset s$, joint observation of $\kappa = [M]$ and $\kappa' = [M']$ requires simply $M \cup M' \subseteq s$, such that $\pi^*_{(\kappa\kappa')} = \Pr(M \cup M' \subseteq s)$ with respect to s^*_{ref}.

5.1.4 Illustration of two strategies using HTE

Using the modified HTE based on $\pi^*_{(\kappa)}$ by (5.4) with respect to $s^*_{\text{ref}} = s \times s$ associated with G^*_s is a different strategy compared to using HTE based on $\pi_{(\kappa)}$ by (5.3) with respect to $s_{\text{ref}} = s \times U \cup U \times s$ associated with G_s. On the one hand, whichever graph total of interest, one may expect a loss of efficiency by using the modified HTE. On the other hand, the HTE requires more computation.

Table 5.1: Inclusion probability of selected triad motif $\kappa = [M]$ in Figure 5.2.

M	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,3,4\}$	$\{2,3,4\}$	$\{1,2,5\}$	$\{1,3,5\}$
$\pi_{(\kappa)}$	0.923	0.853	0.832	0.853	0.867	0.888
$\pi^*_{(\kappa)}$	0.566	0.266	0.203	0.255	0.622	0.538

Table 5.2: Third-order graph total estimate, expectation and standard error

Based on G_s	$\hat{\theta}_{3,0}$	$\hat{\theta}_{3,1}$	$\hat{\theta}_{3,2}$	$\hat{\theta}_{3,3}$
Estimate (Observed)	96.251	89.260	26.289	2.515
Expectation	121	123	40	2
Standard Error	22.977	18.591	7.025	0.768
Based on G^*_s	$\hat{\theta}^*_{3,0}$	$\hat{\theta}^*_{3,1}$	$\hat{\theta}^*_{3,2}$	$\hat{\theta}^*_{3,3}$
Estimate (Observed)	59.128	63.209	19.211	1.607
Expectation	121	123	40	2
Standard Error	78.694	49.929	15.038	1.195

For an illustration, consider 2SBS from the population graph $G = (U, A)$ in Figure 5.2, where $|U| = 13$ and $|A| = 19$, together with the two sample graphs G_s and G^*_s, given the initial sample $s_0 = \{4, 5, 10\}$ by SRS. We have $s_1 = \{1, 2, 8, 9\}$, $s_2 = \{3, 6, 12, 13\}$ and $s = \{1, 2, 4, 5, 8, 9, 10\}$. Table 5.1 lists the inclusion probabilities of 6 selected triad motifs in G and G^*_s, respectively given by (5.3) and (5.4). These can be verified by direct enumeration over all possible initial samples s_0 under SRS with $|s_0| = 3$. Note that every motif in Table 5.1 has a higher inclusion probability in G_s than G^*_s.

Table 5.2 shows the HTE of four 3rd-order graph totals $\hat{\theta}_{3,h}$, for $h = 0, 1, 2, 3$, which are the numbers of triads of size h, as discussed in Section 1.2.4, based on the observed

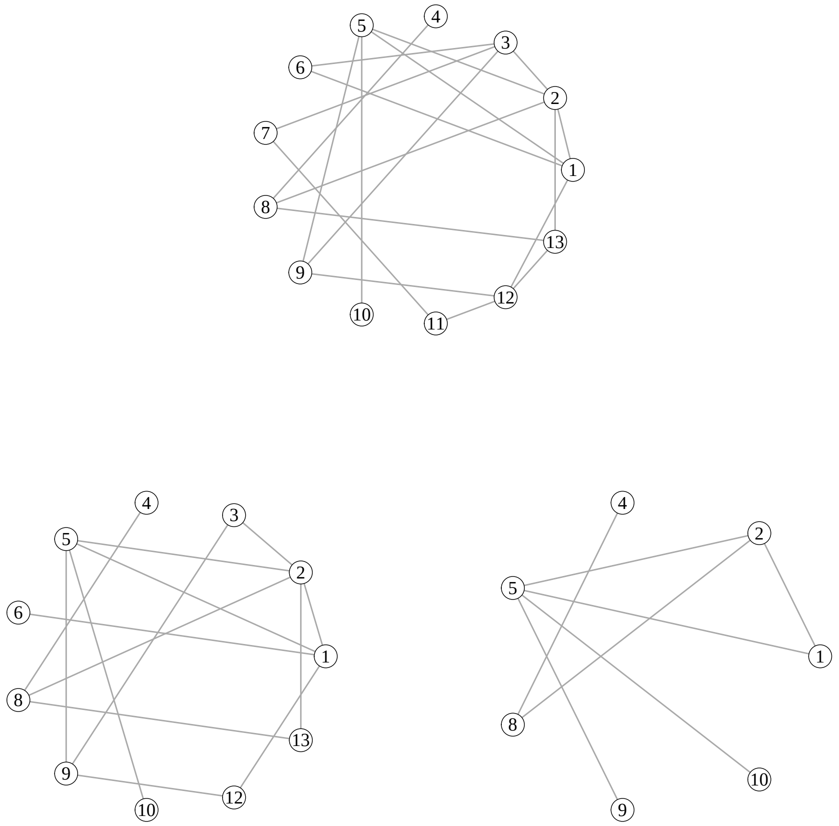

Figure 5.2: Population graph G with 13 nodes and 19 edges (top), sample graphs G_s (bottom left) and G_s^* (bottom right) by 2SBS given initial $s_0 = \{4, 5, 10\}$.

sample graph G_s. Their expectations and standard errors are also given in Table 5.2, which are evaluated by definition, where the expectations are the true totals in G because the HTE is unbiased. Similarly, the modified HTEs based on the observed G_s^* are given in Table 5.2, together with their expectations and standard errors. Clearly, the modified HTE is unbiased but less efficient than the HTE.

5.2 Strategy BIGS-IWE for TSBS

To implement the HTE, one needs to observe the relevant R_L for (5.3), such as R_i for (5.1) or R_{ij} for (5.2). These constitute the ancestry knowledge of the eligible motifs in G_s under TSBS. Similar ancestry knowledge is also needed for the modified HTE.

The circular dependence between ancestry knowledge and eligibility can be handled in two ways essentially. Given the sample motifs Ω_s by TSBS, one can implement additional waves of incident OP if needed, until all the ancestors of each motif in Ω_s under TSBS are observed, so that *all* the motifs in Ω_s become eligible for estimation. One can apply the strategy BIGS-IWE using the BIG given by (3.1), which contains all the observational links from U to Ω under TSBS. Alternatively, without implementing additional waves of OP, one can use *only* the motifs in Ω_s for which the ancestry knowledge is already secured under TSBS. This amounts to limit (3.1) to these identified ancestors, so that the corresponding strategy BIGS-IWE is already applicable, similarly to applying $(\mathcal{B}^*, \hat{\theta}_y)$ or $(\mathcal{B}^\dagger, \hat{\theta}_y)$ for ACS in Sections 4.1.2 and 4.1.3.

For an example, take the 2-star motif of $M = \{1, 2, 3\}$ in the population graph G in Figure 5.2. Starting from $s_0 = \{5\}$, one would observe $\kappa = [M]$ in 2 waves, so that 5 is an ancestor of κ under 2SBS given $|s_0| = 1$. Meanwhile, given $s_0 = \{6\}$, one would also observe this motif, where $s_1 = \{1, 3\}$ and $s_2 = \{2, 5, 7, 9, 12\}$. However, one would not know whether 5 is an ancestor of κ given $s_0 = \{6\}$, since the edge (25) is yet unobserved. One can carry on one more wave, after which (25) would be observed and 5 identified as an ancestor of κ under 2SBS. Alternatively, since one can be sure that $M = \{1, 2, 3\}$ can all be identified as ancestors of κ under 2SBS, one can limit (3.1) to M, so that the corresponding strategy BIGS-IWE is feasible for this motif.

Below we develop theory to substantiate these observations.

5.2.1 Distances to a motif

Let φ_{ij} be the geodesic distance from node i to node j in G. For example, take Figure 5.3, where there is no edge between any of these 7 nodes and the rest of the graph. We have among others $\varphi_{23} = \varphi_{32} = 1$, $\varphi_{12} = \varphi_{13} = \infty$ and $\varphi_{56} = \varphi_{65} = 2$.

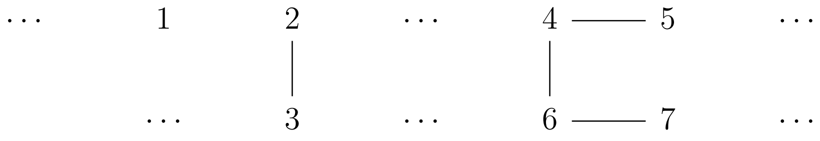

Figure 5.3: Illustration of geodesic distance

The *geodesic distance from node i to motif κ* is the number of waves it takes from i to reach *any* nodes in $M(\kappa)$, denoted by $\varphi_{i,\kappa}$. We have $\varphi_{i,\kappa} = 0$ if $i \in M(\kappa)$, and

$$\varphi_{i,\kappa} = \min_{j \in M(\kappa)} \varphi_{ij}$$

for any $i \notin M(\kappa)$. For instance, take the 2-star motif of $M(\kappa) = \{4, 5, 6\}$ in Figure 5.3, we have $\varphi_{4,\kappa} = \varphi_{5,\kappa} = \varphi_{6,\kappa} = 0$, $\varphi_{7,\kappa} = 1$, and $\varphi_{i,\kappa} = \infty$ for any other node i.

The *radius distance from node i to motif κ* is the number of waves it takes from i to reach *all* the nodes in $M(\kappa)$, denoted by $\lambda_{i,\kappa}$. We have

$$\lambda_{i,\kappa} = \max_{j \in M(\kappa)} \varphi_{ij}$$

for any $i \in U$. Take again $M(\kappa) = \{4, 5, 6\}$ in Figure 5.3, we have e.g. $\lambda_{4,\kappa} = 1$, $\lambda_{5,\kappa} = \lambda_{6,\kappa} = 2$ and $\lambda_{7,\kappa} = 3$, $\lambda_{i,\kappa} = \infty$ for any other node i.

The *observation distance from node i to motif κ* is the number of waves it takes from i to observe the motif κ, denoted by $d_{i,\kappa}$. For instance, take the motif of $M(\kappa) = \{1, 2, 3\}$

in Figure 5.3, we have $d_{1,\kappa} = \infty$ and $d_{2,\kappa} = d_{3,\kappa} = 2$. Whereas, take the motif of $M(\kappa) = \{4, 5, 6\}$, we have $d_{4,\kappa} = d_{5,\kappa} = d_{6,\kappa} = 2$. The calculus of observation distances is summarised in the two following lemmas, for any induced subgraph motif $\kappa = [M]$ which is observed if $M \times M \in s_{\text{ref}}$.

Lemma 5.1. *$\forall \kappa \in \Omega$ and $i \in U$, if the nodes $M(\kappa)$ are connected in G, then*

$$d_{i,\kappa} = \begin{cases} \lambda_{i,\kappa} & \text{if } |\arg\max_{j\in M(\kappa)} \varphi_{ij}| = 1 \\ 1 + \lambda_{i,\kappa} & \text{otherwise} \end{cases}$$

Proof. If $|\arg\max_{j\in M_\kappa} \varphi_{ij}| = 1$, then there is only one node in $M(\kappa)$, denoted by j_0, which requires $\lambda_{i,\kappa}$ waves from i. All the other nodes in $M(\kappa)$ must be observed before j_0, which allows one to observe any edge between them at the latest by the last wave, when j_0 is observed. If there are more than one node that requires $\lambda_{i,\kappa}$ waves from i, then an additional wave is needed to observe the edges among them. □

Lemma 5.2. *$\forall \kappa \in \Omega$ and $i \in U$, if there exists a single node in $M(\kappa)$ which is unconnected to i in G, then*

$$d_{i,\kappa} = 1 + \max_{j\in M(\kappa;i)} \varphi_{ij}$$

where $M(\kappa; i)$ consists of the nodes in $M(\kappa)$ that are connect to i in G.

Proof. All the nodes $M(\kappa; i)$ are reached from i by $\max_{j\in M(\kappa;i)} \varphi_{ij}$ waves and not earlier. An additional wave is needed to confirm that the last node is unconnected to $M(\kappa; i)$, as well as to ensure that one observes all the edges among $M(\kappa; i)$. □

Corollary 5.1. *If there are at least two nodes $i, j \in M(\kappa)$ with $d_{i,\kappa} = d_{j,\kappa} = \infty$, then the strategy BIGS-IWE is inapplicable to κ.*

Proof. One cannot observe κ by incident OP starting from any node h in U because, by stipulation, h can possibly be connected to either i or j but not both. Hence, no node in F can satisfy (3.1). □

5.2.2 Using β_κ by (3.1)

The node i in G is a *TSBS ancestor* of motif κ if $d_{i,\kappa} \le T$. Any TSBS ancestor i of given motif κ satisfies (3.1) by definition. Let

$$\varphi_\kappa = \max_{i,j\in M(\kappa)} \varphi_{ij}$$

be the *diameter* of the motif κ. Let

$$\zeta_\kappa = \max_{i\in M(\kappa)} d_{i,\kappa}$$

be the *observation diameter* of the motif κ. By Lemma 5.1, given any connected $M(\kappa)$ with $\varphi_\kappa < \infty$, we have

$$\zeta_\kappa \le 1 + \varphi_\kappa$$

Lemma 5.3. *$\forall \kappa \in \Omega_s$ with $\varphi_\kappa < \infty$, if $|M(\kappa)| > 1$ then one needs at most $T-1$ waves of incident OP from $M(\kappa)$ to observe all the ancestors of κ under TSBS from G, if $|M(\kappa)| = 1$ then one needs at most T waves of incident OP from $M(\kappa)$.*

Proof. Applying $T-1$ waves of incident OP to $M(\kappa)$, as if $s_0 = M(\kappa)$, would identify *all* the nodes $\{i : \varphi_{i,\kappa} \leq T-1, i \in U\}$. If $|M(\kappa)| > 1$, then only a subset of them can be all the nodes $\{i : \lambda_{i,\kappa} \leq T, i \in U\}$. By Lemma 5.1, only a subset of them can be all the nodes $\{i : d_{i,\kappa} \leq T, i \in U\}$. If $|M(\kappa)| = 1$, then applying at most T waves of incident OP to $M(\kappa)$ would identify *all* the nodes $\{i : d_{i,\kappa} \leq T, i \in U\}$. □

Having identified all the TSBS ancestors of each κ in Ω_s by Lemma 5.3, one obtains β_κ under BIGS from $\mathcal{B}$ given by (3.1), by which the condition (ii) of Theorem 3.1 is satisfied. The condition (i) is satisfied if every motif κ in Ω has a positive inclusion probability in Ω_s under TSBS from G. The strategy BIGS-IWE yields then unbiased estimation of any graph total θ defined over Ω.

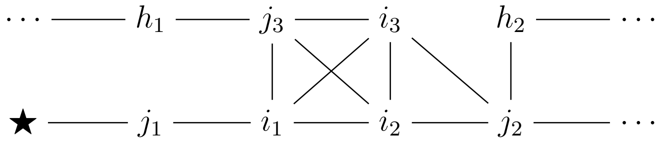

Figure 5.4: Illustration of strategy BIGS-IWE for TSBS

Notice that Lemma 5.3 provides an upper bound for the number of additional waves. For instance, consider 3SBS and the triangle κ with nodes $M(\kappa) = \{i_1, i_2, i_3\}$ in Figure 5.4. We have $\varphi_{i,\kappa} = 1$ for $i \in \{j_1, j_2, j_3\}$, $\lambda_{j_3,\kappa} = 1$ and $\lambda_{i,\kappa} = 2$ for $i \in \{j_1, j_2\}$; moreover, $\varphi_{i,\kappa} = 2$ for $i \in \{h_1, h_2\}$, $\lambda_{h_1,\kappa} = 2$ and $\lambda_{h_2,\kappa} = 3$. The 3SBS ancestors of κ are $\{i_1, i_2, i_3, j_2, j_3\}$ with $d_{i,\kappa} = 2$ and $\{j_1, h_1, h_2\}$ with $d_{i,\kappa} = 3$. Given $\kappa \in \Omega_s$, we need at most 2 more waves to identify all its 3SBS ancestors by Lemma 5.3. Indeed, in this case we only need at most 1 more wave to identify all the 3SBS ancestors from the moment κ is observed, which can be verified by enumerating all the possibilities when *only* one 3SBS ancestor is selected in s_0. For instance, we need one more wave to observe h_2 starting from only $h_1 \in s_0$. Another wave from h_2 is unnecessary because $d_{h_2,\kappa} = 3$ and h_2 is not adjacent to any observed 2SBS ancestors except from j_2.

5.2.3 Using subset of β_κ

By a similar treatment of ACS in Section 4.1.2, one can limit (3.1) to a subset of TSBS ancestors which can be determined *a priori*, without additional waves of OP.

For each $\kappa \in \Omega$, let $\mathcal{B}$ be the population BIG constructed by (3.1), where β_κ consists of all the TSBS ancestors of κ. Let β_κ^* be a non-empty subset of β_κ, where $\emptyset \neq \beta_\kappa^* \subseteq \beta_\kappa$. Let H^* contain the edges from β_κ^* to κ, for each $\kappa \in \Omega$. Let

$$\mathcal{B}^* = (F, \Omega; H^*) \qquad \text{and} \qquad H^* = \bigcup_{\kappa \in \Omega} \beta_\kappa^* \times \kappa \tag{5.5}$$

Denote by ($\mathcal{B}^*$, IWE) the BIGS-IWE strategy, where a sample motif κ in Ω_s (by TSBS) is eligible for IWE iff $s_0 \cap \beta_\kappa^* \neq \emptyset$.

Since β_κ^* is non-empty for every $\kappa \in \Omega$, the condition (i) of Theorem 3.1 remains satisfied. The condition (ii) of Theorem 3.1 is satisfied if β_κ^* can be identified as TSBS ancestors based on G_s under TSBS, *whenever* $\beta_\kappa^* \cap s_0 \neq \emptyset$. Two results below can ensure that this is the case under the given TSBS, which means that the strategy ($\mathcal{B}^*$, IWE) can then be determined without taking any sample graph.

Lemma 5.4. *Provided $\zeta_\kappa < \infty$ of all $\kappa \in \Omega$, the condition (ii) of Theorem 3.1 is satisfied given* (5.5)*, where $\beta_\kappa^* = M(\kappa)$, for TSBS with $T = \max_{\kappa\in\Omega} \zeta_\kappa$.*

Proof. Given $T = \max_{\kappa\in\Omega} \zeta_\kappa$, a given motif κ and all the edges in $M(\kappa) \times M(\kappa)$ are observed under TSBS if $s_0 \cap M(\kappa) \neq \emptyset$, such that $\beta_\kappa^* = M(\kappa)$ can be identified as TSBS ancestors based on the corresponding sample graph G_s. □

Additional waves of observation is not needed for applying the IWE under BIGS from $\mathcal{B}^*$ given by (5.5) and Lemma 5.4. But fewer sample motifs may be eligible compared to using BIGS from $\mathcal{B}$ containing all the TSBS ancestors, which generally requires additional waves of OP. When the uncertainty associated with ($\mathcal{B}^*$, IWE) is too large given $T = \max_{\kappa\in\Omega} \zeta_\kappa$, one may extend the SBS in order to obtain a larger sample graph. This raises the need to update the BIG for TSBS, where $T > \max_{\kappa\in\Omega} \zeta_\kappa$. Let

$$\beta_\kappa^t(M) = \{i \notin M(\kappa) : \varphi_{i,\kappa} \leq t\}$$

contain all the nodes *outside of* $M(\kappa)$, which have maximum geodesic distance t to κ. That is, starting from any $i \in \beta_\kappa^t(M)$, it takes at most t waves of incident OP in G to reach $M(\kappa)$. Given $T > \max_{\kappa\in\Omega} \zeta_\kappa$, the nodes in $\beta_\kappa^t(M)$ may be accepted to β_κ^*, for $t \geq 1$, according to the result below.

Lemma 5.5. *Provided $\zeta_\kappa < \infty$ of all $\kappa \in \Omega$, the condition (ii) of Theorem 3.1 is satisfied given* (5.5)*, where $\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^t(M)$, for TSBS with $T = \max_{\kappa\in\Omega} \varphi_\kappa + 2t$ and $t \geq 1$.*

Proof. Any given motif κ is observed after at most $\zeta_\kappa + t$ waves starting from any node in β_κ^* defined above. First, if $\zeta_\kappa = 1 + \varphi_\kappa$, then all the nodes $M(\kappa)$ must have been observed at wave $\zeta_\kappa + t - 1$, so that all the nodes $\beta_\kappa^1(M)$ are already observed after $\zeta_\kappa + t = \varphi_\kappa + t + 1$ waves. It remains only to observe all the geodesics to $\beta_\kappa^t(M) \setminus \beta_\kappa^1(M)$ starting from $\beta_\kappa^1(M)$, which requires at most $t-1$ waves. Next, if $\zeta_\kappa = \lambda_\kappa$, then there is at least one node $j \in M(\kappa)$, which is first observed at wave ζ_κ starting from any node in $M(\kappa)$. Up to t additional waves may be needed to observe all the nodes outside $M(\kappa)$, which can lead to j in t waves. Thus, in either case, β_κ^* can be identified as TSBS ancestors based on the observed sample graph G_s. □

Take the triangle κ with $M(\kappa) = \{i_1, i_2, i_3\}$ in Figure 5.4. We have $\varphi_\kappa = \zeta_\kappa = 2$. By Lemma 5.4, one can let $\beta_\kappa^* = M(\kappa)$ for 2SBS, which excludes $\{j_2, j_3\}$ that are also 2SBS ancestors. By Lemma 5.5, setting $\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^1(M)$ is feasible for 4SBS, where $\beta_\kappa^1(M) = \{j_1, j_2, j_3\}$. One can enumerate all possible sample graphs G_s given $s_0 \cap \beta_\kappa^* \neq \emptyset$, to verify that β_κ^* can always be identified as 4SBS ancestors based on G_s. Notice that Lemma 5.5 does not permit one to include $\{h_1, h_2\}$ in β_κ^*, which are 3SBS ancestors of κ, because the construction of β_κ^* is based on $\beta_\kappa^t(M)$ defined in terms of $\varphi_{i,\kappa}$. For instance, the ★-node has the same geodesic distance to κ as $\{h_1, h_2\}$ in Figure 5.4. However, starting from only $★ \in s_0$, we would observe κ under 4SBS but not (j_2h_2), such that h_2 can not be identified as a 4SBS ancestor on this occasion.

5.2.4 Using subset of $\beta_{\kappa|s}$

By a similar treatment of ACS in Sections 4.1.3 and 4.1.4, one can limit (3.1) to a subset of sample TSBS ancestors without additional waves of observation, and make inference conditional on the chosen strategy.

Given the sample graph G_s observed under TSBS from G, let $d_{i,\kappa}(G_s)$ be the *sample observation distance* from i to given motif κ in G_s, where $d_{i,\kappa}(G_s) \geq d_{i,\kappa}(G)$ generally. Given any $\kappa \in \Omega_s$, the observed sample TSBS ancestors are

$$\beta_{\kappa|s} = \{i : d_{i,\kappa}(G_s) \leq T\} \tag{5.6}$$

Note that for each $i \in \beta_{\kappa|s}$, there exists at least a path corresponding to $d_{i,\kappa}(G_s)$ in the sample subgraph induced by $\beta_{\kappa|s}$, denoted by $G_s(\beta_{\kappa|s})$. This is because any node on the path giving rise to $d_{i,\kappa}(G_s)$ must have a shorter sample observation distance to κ than i, so that it must belong to $\beta_{\kappa|s}$ as well.

Note that it may not be possible to observe the whole $\beta_{\kappa|s}$ whenever $\beta_{\kappa|s} \cap s_0 \neq \emptyset$. Let $\beta_{\kappa|s}$ form an *ancestor network* of κ if the subgraph $G_s(\beta_{\kappa|s})$ induced by it can always be observed whenever $s_0 \cap \beta_{\kappa|s} \neq \emptyset$. In other words, $\beta_{\kappa|s}$ given by (5.6) is an ancestor network of κ if for any $i \in \beta_{\kappa|s}$, we have

$$d_{i,\tau}(G_s) \leq T \quad \text{where} \quad \tau = [G_s(\beta_{\kappa|s})] \tag{5.7}$$

Of course, in case not all the nodes in $\beta_{\kappa|s}$ by (5.6) satisfy the condition (5.7), it is always possible to find a subset $\beta^*_{\kappa|s}$ of $\beta_{\kappa|s}$, which form an ancestor network, where

$$\beta^*_{\kappa|s} = \{i : d_{i,\kappa}(G_s) \leq T,\ d_{i,\tau}(G_s) \leq T,\ \tau = [G_s(\beta^*_{\kappa|s})]\} \tag{5.8}$$

By Theorem 4.1, one can chose an fixed non-empty subset β'_κ of $\beta_{\kappa|s}$, including $\beta_{\kappa|s}$ itself or an ancestor network $\beta^*_{\kappa|s}$ derived from it. The BIGS-IWE strategy using the corresponding $\mathcal{B}' = (F, \Omega; H')$ is unbiased for θ, and the inference is conditional on the chosen strategy. This sample-dependent approach is often more efficient than the strategy by Lemma 5.5, because one tends to have $\beta^*_\kappa \subset \beta^*_{\kappa|s}$.

For an illustration, consider the triangle motif of $M(\kappa) = \{i_1, i_2, i_3\}$ in Figure 5.4. First, for 2SBS, we have $\beta^*_\kappa = M(\kappa)$ by Lemma 5.4 or Lemma 5.5, whereas $\beta_{\kappa|s} = \beta^*_{\kappa|s} = M(\kappa) \cup \{j_2, j_3\}$ by Theorem 4.1 which are all the 2SBS ancestors of κ here.

Next, Lemma 5.5 yields $\beta^*_\kappa = M(\kappa)$ for 3SBS, which is ineffective because it is the same as for 2SBS. The 3SBS ancestors are $\beta_\kappa = M(\kappa) \cup \{j_2, j_3\} \cup \{j_1, h_1, h_2\}$.

- If $s_0 \cap \beta_\kappa = \{h_2\}$, then neither (i_1j_1) nor (j_3h_1) is observed, such that

$$\beta_{\kappa|s}(h_2) = \beta^*_{\kappa|s}(h_2) = M(\kappa) \cup \{j_2, j_3\} \cup \{h_2\}$$

Notice that $d_{j_3,\tau}(G_s) = 2$, although $(i_1 j_3)$ is yet unobserved.

- If $s_0 \cap \beta_\kappa = \{j_1\}$ or $\{h_1\}$, then $(j_2 h_2)$ is unobserved, such that

$$\beta_{\kappa|s}(j_1, h_1) = \beta^*_{\kappa|s}(j_1, h_1) = M(\kappa) \cup \{j_2, j_3\} \cup \{j_1, h_1\}$$

- In all the other situations where κ is observed under 3SBS, one would observe

$$\beta_{\kappa|s} = M(\kappa) \cup \{j_2, j_3\} \cup \{j_1, h_1, h_2\} = \beta_\kappa$$

which does not satisfy (5.7). For an ancestor network, one can let $\beta^*_{\kappa|s} = \beta^*_{\kappa|s}(h_2)$ by removing $\{j_1, h_1\}$ or let $\beta^*_{\kappa|s} = \beta^*_{\kappa|s}(j_1, h_1)$ by removing h_2.

Finally, for 4SBS, Lemma 5.5 yields $\beta^*_\kappa = M(\kappa) \cup \{j_1, j_2, j_3\}$. In addition to the 3SBS ancestors, β_κ now includes also the nodes like ★. We notice the following.

- If $\beta^*_{\kappa|s}$ by (5.8) includes the ★-node in Figure 5.4, then it must exclude h_2 and the other nodes adjacent to j_2 or h_1.
- If $\beta^*_{\kappa|s}$ by (5.8) includes h_2 and other nodes adjacent to j_2, then it must exclude the ★-node and the other nodes adjacent to h_1.
- If $M(\kappa) \cup \{j_1, j_2, j_3, h_1\}$ intersects s_0, then $\beta_{\kappa|s}$ would include h_2 and ★, as well as the nodes adjacent to j_2 and h_1, although not all of them can be included in $\beta^*_{\kappa|s}$.

Anyway, we always have $\beta^*_\kappa \subset \beta^*_{\kappa|s}$ here, so that a sample-dependent strategy following Theorem 4.1 can be more efficient than one following Lemma 5.5.

Generally, different ancestor networks $\beta^*_{\kappa|s}$ can be observed across all possible sample graphs where κ is observed. It is potentially possible to identify a subset of them, which is common to all these sample graphs. However, doing so is not worthwhile if it cannot improve the efficiency. For instance, under 3SBS from Figure 5.4, we have $\beta^*_{\kappa|s}(h_2) \cap \beta^*_{\kappa|s}(j_1, h_1) = M(\kappa) \cup \{j_2, j_3\}$, but there is no advantage in adopting it instead of either $\beta^*_{\kappa|s}(h_2)$ or $\beta^*_{\kappa|s}(j_1, h_1)$ depending on the actual sample graph.

5.3 Illustration of BIGS-IWE for SBS

Figure 5.5 shows a population graph G of 40 nodes and 72 edges. Consider the motifs illustrated in Figure 1.2: node ($\mathcal{K}_1$), 2-clique ($\mathcal{K}_2$), 2-star ($\mathcal{S}_2$), 3-clique ($\mathcal{K}_3$), 4-clique ($\mathcal{K}_4$), 4-cycle ($\mathcal{C}_4$), 3-star ($\mathcal{S}_3$) and 3-path ($\mathcal{P}_3$). Their totals in Figure 5.5 are

$$(\theta_{\mathcal{K}_1}, \theta_{\mathcal{K}_2}, \theta_{\mathcal{S}_2}, \theta_{\mathcal{K}_3}, \theta_{\mathcal{K}_4}, \theta_{\mathcal{C}_4}, \theta_{\mathcal{S}_3}, \theta_{\mathcal{P}_3}) = (40, 179, 72, 19, 3, 7, 141, 408)$$

The diameters φ_κ and the observation diameters ζ_κ of the motifs are given at the top of Table 5.3. For an IWE such as the HTE $\hat{\theta}_y$ and the multiplicity estimator $\hat{\theta}_{z\beta}$ under BIGS from (5.5) given $\beta^*_\kappa = M(\kappa)$, ζ_κ waves of SBS is required by Lemma 5.4, whereas one may need up to ζ_κ additional waves for the HH-type estimator $\hat{\theta}_{z\alpha 1}$ using the PIDA

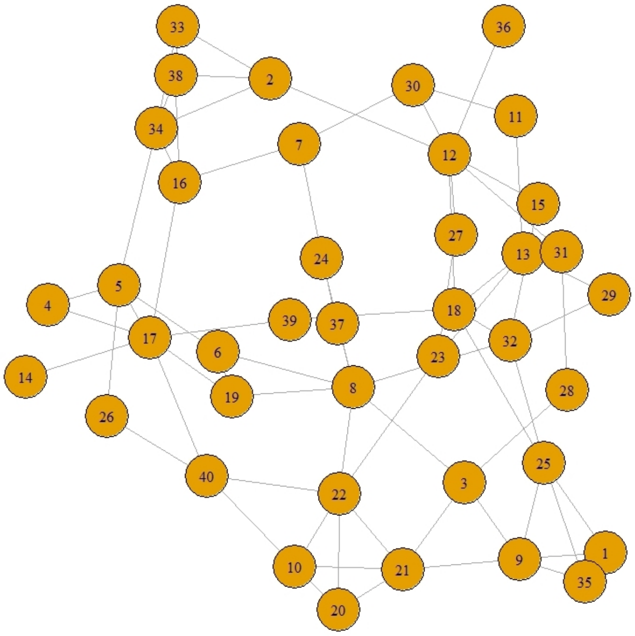

Figure 5.5: A population graph with $|U| = 40$ and $|A| = 72$.

Table 5.3: Diameter and observation diameter of motifs. Number of waves T required for applying selected IWE under BIGS from (5.5) given specified β^*_κ.

		$\mathcal{K}_1$	$\mathcal{K}_2$	$\mathcal{S}_2$	$\mathcal{K}_3$	$\mathcal{K}_4$	$\mathcal{C}_4$	$\mathcal{S}_3$	$\mathcal{P}_3$
	φ_κ	0	1	2	1	1	2	2	3
β^*_κ	ζ_κ	0	1	2	2	2	2	3	3
$M(\kappa)$	T for $\hat{\theta}_{z\beta}$	0	1	2	2	2	2	3	3
	T for $\hat{\theta}_{z\alpha 1}$	0	2	4	3	3	4	5	6
$M(\kappa) \cup \beta^1_\kappa(M)$	T for $\hat{\theta}_{z\beta}$	2	3	4	3	3	4	4	5
	T for $\hat{\theta}_{z\alpha 1}$	3	5	7	6	6	7	8	9
$M(\kappa) \cup \beta^2_\kappa(M)$	T for $\hat{\theta}_{z\beta}$	4	5	6	5	5	6	6	7
	T for $\hat{\theta}_{z\alpha 1}$	6	8	10	9	9	10	11	12

weights (2.9) with $\gamma = 1$. Next, $\varphi_\kappa + 2t$ waves of SBS is required by Lemma 5.5 for $\hat{\theta}_y$ and $\hat{\theta}_{z\beta}$ under BIGS from (5.5) given $\beta^*_\kappa = M(\kappa) \cup \beta^t_\kappa(M)$ with $t \geq 1$, whereas up to $\zeta_\kappa + t$ additional waves of observation may be needed for $\hat{\theta}_{z\alpha 1}$.

5.3.1 An example: $s_0 = \{3, 12\}$ and $[M] = \mathcal{C}_4$

To illustrate some of the computational details, let 4-cycle $\mathcal{C}_4$ be the motif of interest. The graph total of $\mathcal{C}_4$ is 7. In Figure 5.6, the initial sample $s_0 = \{3, 12\}$ by SRS is marked as TSBS with $T = 0$; the sample graphs observed under TSBS are marked as $T = 1, 2, 3, 4$. The sample graph by 4SBS includes all the nodes in G but not all the edges.

According to Lemma 5.4, using $\beta^*_\kappa = M(\kappa)$ is feasible for 2SBS for $\mathcal{C}_4$, where $\zeta_\kappa = 2$ (Table 5.3). The details required for computing $\hat{\theta}_y$, $\hat{\theta}_{z\beta}$ and $\hat{\theta}_{z\alpha 1}$ are given in the upper part of Table 5.4, where $\beta^*_\kappa = M(\kappa)$. The motif A with nodes $M(A) = \{3, 8, 21, 22\}$ is observed

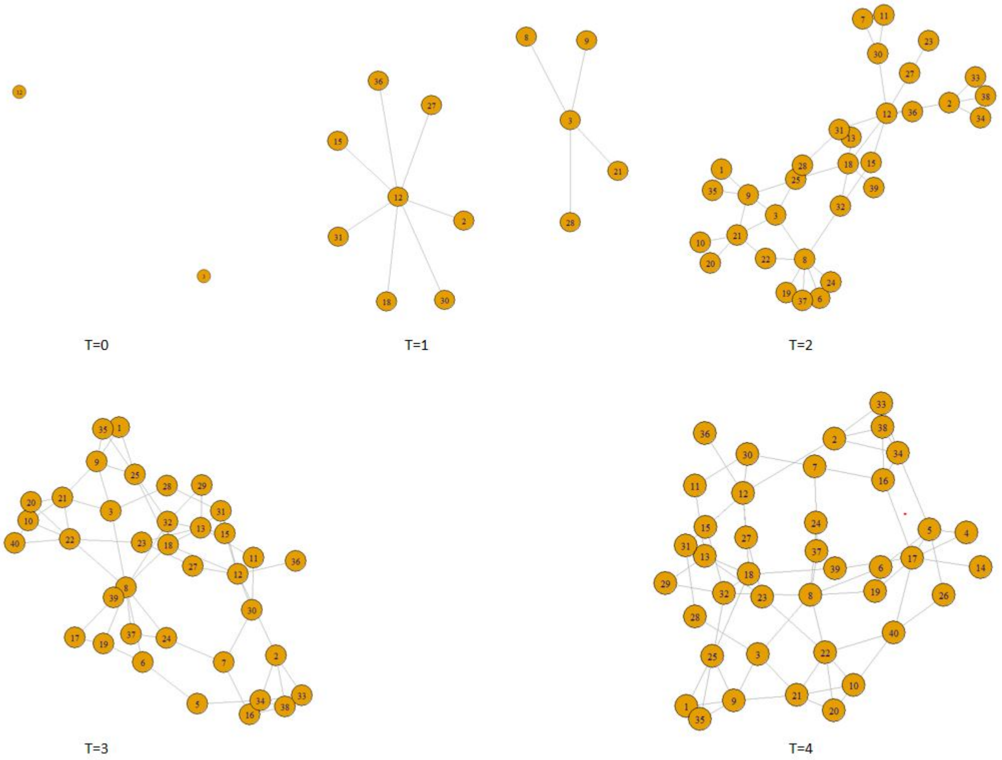

Figure 5.6: Initial sample $s_0 = \{3, 12\}$, sample graphs by TSBS for $T \leq 4$.

from node 3, and the motifs B and C from node 12, where $M(B) = \{12, 13, 18, 31\}$ and $M(C) = \{12, 15, 18, 32\}$. Two more waves are needed to apply $\hat{\theta}_{z\alpha 1}$, where the relevant $|\alpha_i^*|$ in $\mathcal{B}^*$ are given in the last column of Table 5.4. For instance, we have $\{|\alpha_i^*| : i \in \beta_B^*\} = \{|\alpha_{12}^*|, |\alpha_{13}^*|, |\alpha_{18}^*|, |\alpha_{31}^*|\} = \{2, 2, 3, 1\}$ for the motif B.

Table 5.4: Applying BIGS-IWE for $\mathcal{C}_4$ under TSBS given $s_0 = \{3, 12\}$.

	$i \in s_0$	$\kappa \in \alpha_i^*$	$M(\kappa)$	$\lvert\beta_\kappa^*\rvert$	$\{\lvert\alpha_i^*\rvert : i \in \beta_\kappa^*\}$
$T = 2$	3	A	$\{3, 8, 21, 22\}$	4	$\{1, 1, 1, 1\}$
$\beta_\kappa^* = M(\kappa)$	12	B	$\{12, 13, 18, 31\}$	4	$\{2, 2, 3, 1\}$
		C	$\{12, 15, 18, 32\}$	4	$\{2, 1, 3, 2\}$
$T = 4$	3	A	$\{3, 8, 21, 22\}$	15	–
$\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^1(M)$	12	B	$\{12, 13, 18, 31\}$	16	–
		C	$\{12, 15, 18, 32\}$	14	–
		D	$\{13, 18, 29, 32\}$	12	–

Under SRS of s_0 with $m = |s_0|$, the inclusion probability of a motif κ is

$$\pi_{(\kappa)} = 1 - \binom{N - |\beta_\kappa^*|}{m} / \binom{N}{m} \equiv 0.1923$$

where $N = 40$, $m = 2$ and $|\beta_\kappa^*| \equiv 4$ for $\mathcal{C}_4$. By (2.6), we have $\hat{\theta}_y = 3/0.1923 = 15.6$. For $\hat{\theta}_{z\beta}$ by (2.7), we have $z_3 = 1/4$ from $\alpha_3^* = \{A\}$ and $z_{12} = 1/4 + 1/4$ from $\alpha_{12}^* = \{B, C\}$, such that $\hat{\theta}_{z\beta} = (3/4)/(2/40) = 15$, where $\pi_i \equiv 2/40$. For $\hat{\theta}_{z\alpha 1}$ by (2.7), we have $z_3 = 1/4$ from α_3^* and $z_{12} = 0.5/2.33 + 0.5/2.33 = 0.43$ from α_{12}^*, such that $\hat{\theta}_{z\alpha 1} = 13.6$.

By Lemma 5.5, setting $\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^1(M))$ with $t = 1$ is feasible for $\mathcal{C}_4$ under 4SBS, where $\varphi_\kappa = 2$ (Table 5.3). The details are given in the lower part of Table 5.4. The motif A is observed from node 3, and the motifs $\{B, C, D\}$ from node 12 where, compared to 2SBS above, the extra motif D with $M(D) = \{13, 18, 29, 32\}$ is observed via node 18 obtained at the 1st wave (Figure 5.6). All these motifs are observed by the 3rd wave; however, since $\varphi_\kappa = \zeta_\kappa$ for motif $\mathcal{C}_4$, another wave is needed to ensure the ancestry knowledge, yielding $T = \varphi_\kappa + 2t = 4$. The cardinality of the ancestor set β_κ^* is given in Table 5.4, which is 15, 16, 14, or 12 for $\kappa = A, B, C$ or D. More waves of observation are needed to apply $\hat{\theta}_{z\alpha 1}$, so that it is not feasible under 4SBS illustrated here, and the relevant $|\alpha_i^*|$ are omitted in Table 5.4.

The inclusion probability $\pi_{(\kappa)}$ is now 0.6154, 0.6462, 0.5833 and 0.5154 for $\kappa = A, B, C$ and D, which are much higher than $\pi_{(k)} \equiv 0.1923$ under 2SBS above. By (2.6), we have $\hat{\theta}_y = 6.83$. For $\hat{\theta}_{z\beta}$ by (2.7), we have $z_3 = 1/15$ from $\alpha_3^* = \{A\}$ and $z_{12} = 1/16+1/14+1/12$ from $\alpha_{12}^* = \{B, C, D\}$, such that $\hat{\theta}_{z\beta} = 5.68$. Both these two estimates are much closer to the graph total $\theta_{\mathcal{C}_4} = 7$ than those under 2SBS, with only one extra motif D. Contrasting SBS with $T = 4$ or $T = 2$, the inclusion probabilities $\pi_{(\kappa)}$ and the weights $\omega_{i\kappa}$ matter more to estimation than the number of eligible sample motifs.

5.3.2 Results

Consider SBS up to 4 waves given SRS of s_0 with $|s_0| = 2$. Since the diameter of the population graph G is 6 here, a large part of it may be observed under 4SBS, as in the case of $s_0 = \{3, 12\}$ above; indeed, G is fully observed for 215 out of 780 possible initial samples. In addition, we consider induced OP following SRS of s, for which $s_{\text{ref}} = s \times s$, where $|s|$ 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 HTE.

Table 5.5: Mean squared errors of graph total estimators under induced OP from SRS of size 9 or 21, and SBS of maximum 4 waves given 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, $\lvert s\rvert = 9$	$\hat{\theta}$	1 263	47 134	2 869	2 167	5 168	231 805	797 578
Induced OP, $\lvert s\rvert = 21$	$\hat{\theta}$	152	4 533	198	41	116	11 523	52 488
$\beta_\kappa^* = M(\kappa)$	$\hat{\theta}_y$	471	5 269	193	10	38	5 092	27 717
	$\hat{\theta}_{z\beta}$	475	5 447	199	10	39	5 368	29 441
	$\hat{\theta}_{z\alpha 1}$	116	613	160	10	28	–	–
$\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^1(M)$	$\hat{\theta}_y$	306	1 614	92	4	7	1 382	–
	$\hat{\theta}_{z\beta}$	281	1 485	98	5	7	1 403	–

Table 5.5 gives the mean squared errors of the different estimators. The strategy BIGS-IWE is applied for TSBS. In case an estimator is not feasible for a certain motif under 4SBS, the result is unavailable in the table.

Induced OP is understandably much less efficient than incident OP, as the order of the motif increases; compare e.g. the results for SRS of size 21 and 2SBS using $\beta_\kappa^* = M(\kappa)$, where both have the same expected number of nodes in the sample graph.

Under TSBS from the population graph in Figure 5.5, the HTE $\hat{\theta}_y$ and the HH-type estimator $\hat{\theta}_{z\beta}$ are about equally efficient for the motifs considered here. The HH-type estimator $\hat{\theta}_{z\alpha 1}$ 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 (2.7) is minimised, if the constructed z_i happens to be constant across $i \in F$. Based on equal weights, z_i is proportional to $|\alpha_i|$. Setting $\omega_{ik} \propto |\alpha_i|^{-1}$ tends to even out the values of z_i, since a node with many successors will receive relatively little share from each motif observed from it, when it is based on more motifs than another node with fewer successors.

Bibliographic notes

Goodman (1961) considers snowball sampling on a special directed graph, where each node has one and only one out-edge. Frank (1977a) and Frank and Snijders (1994) consider one-wave SBS from arbitrary population graphs.

Zhang and Patone (2017) derive the HT-estimator for T-wave snowball sampling in general, including the illustrations in Section 5.1.

The incident OP we have considered includes all the successors α_i. It is possible to take only some of the successors at each wave (e.g. Snijders, 1992). In particular, taking randomly one successor each time yields a random walk (e.g. Klovdahl, 1989). However, the ancestry knowledge cannot be ensured under subsampling from α_i at each wave. A different approach to graph sampling strategy is needed, as discussed in Chapter 6.

Frank (1971) defines the *reach* at i as the order of the connected component containing it. Provided one is able to observe the reach, without actually sampling the whole connected component, the ancestry knowledge would be readily available. Whether such an OP is applicable depends on the conditions one operates in.

The approach of BIGS-IWE for TSBS based on Lemmas 5.4 and 5.5, as well as the numerical illustration in Section 5.3, are developed by Zhang and Oguz-Alper (2020).

Chapter 6

Targeted random walk sampling

Random walk can be regarded as a probabilistic depth-first search algorithm in graphs, where the initial node does not need to have a known selection probability, which can be attractive for large and often dynamic graphs if the walk is fast-moving. Although the sample graph inclusion probability is intractable, the successive sequential sampling probability at equilibrium can provide the basis of inference.

6.1 Random walk in graphs

6.1.1 Random walk

Let $G = (U, A)$ be a simple graph. Let $X_t = i$ be the node (or *state*) at step time t. Let a_{i+} be the number of out-edges from node i. For time $t + 1$, one selects one of the out-edges randomly, $(ij) \in A_{i+}$, which yields $X_{t+1} = j$ as the next state of the random walk. Thus, $\{X_0, X_1, X_2, ...\}$ form a Markov chain, which is a basic type of stochastic process, where X_0 is the initial state and the transition probability is

$$p_{ij} := \Pr(X_{t+1} = j | X_t = i) = \frac{a_{ij}}{a_{i+}}$$

Let P be the $N \times N$ matrix of *transition probabilities* with elements p_{ij}. Let p_0 be the *row* N-vector of initial node selection probabilities, the probabilities of X_t are given by

$$p_t = p_0 P^t$$

A random walk reaches gradually its *equilibrium*, if the chance that it visits a given node depends less and less on the initial state X_0. The *stationary probability* of $X_t = i$ is the fraction of times node i is visited when the walk is at equilibrium, denoted by

$$\pi_i = \Pr(X_t = i) \quad \text{and} \quad \pi = (\pi_1, ..., \pi_N)$$

where $\sum_{i \in U} \pi_i = 1$. If G consists of a single component, then every node can be reached from any other node in time and the chain $\{X_t\}$ is irreducible. We have

$$\pi_j = \sum_{j \in U} \pi_i p_{ij} \qquad \text{or} \qquad \pi = \pi P \tag{6.1}$$

where π is the left eigenvector of P with eigenvalue 1. For an undirected graph, where $a_{ij} = a_{ji}$ and $a_{i+} = d_i$ is the degree of node i, we have the main result

$$\pi_i = \frac{d_i}{2R} \tag{6.2}$$

where $R = |A|$. Moreover, for two adjacent nodes i and j, we have

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{6.3}$$

i.e. the flow probability in either direction of each edge is the same at equilibrium. For digraphs, the stationary probabilities satisfying (6.1) cannot be given explicitly.

Only by walking one cannot move beyond the component, to which the initial node belongs. Random jumps can be introduced generally. At each step time, if possible let the walk move to an adjacent node with probability r or jump to *any* node in U with a probability $1 - r$. The transition probabilities are then given by

$$p_{ij} = \begin{cases} r\frac{a_{ij}}{a_{i+}} + (1-r)\frac{1}{N} & \text{if } a_{i+} > 0 \\ \frac{1}{N} & \text{if } a_{i+} = 0 \end{cases} \tag{6.4}$$

We have

$$p_{t+1} = r p_t P + (1-r)u \quad \text{with } N\text{-vector} \quad u = \Big(\frac{1}{N}, \cdots, \frac{1}{N}\Big)$$

From now on, by random walk we include random jumps, unless otherwise specified.

Adding random jumps ensures the chain $\{X_t\}$ is irreducible. Since it is possible to return to the same node at any step time, the chain is aperiodic. The limiting probabilities are the stationary probabilities, which are given by

$$\pi = (1-r)u(I - rP)^{-1}$$

The dependence on r can be seen from the Taylor expansion, which is given by

$$\pi_i \approx \frac{1}{N} + \sum_{l=1}^{\infty} \frac{r^l}{N} \sum_{j=1}^{N} \big\{ \big(\frac{a_{ji}}{a_{j+}}\big)^l - \big(\frac{a_{ji}}{a_{j+}}\big)^{l-1} \big\}$$

However, π remains unknown, without observing the whole matrix P. One can only calculate π for the observed sample graph as if it were the whole graph.

6.1.2 Examples of related topics

PageRank The stationary probability π_i of random walk with random jumps provides a graph centrality measure of the node. One can allow u_i to differ for the nodes in U, in which case u is referred to as the preference vector. The corresponding stationary π can e.g. be used to rank web pages, where it is known as PageRank, where the constant r in (6.4) is called the "damping factor" (Brin and Page, 1998). Given $u_i = a_{+i}/\sum_{j\in U} a_{+j}$, the Taylor expansion becomes

$$\pi_i \approx \frac{a_{+i}}{\sum_{j\in U} a_{+j}} + \sum_{l=1}^{\infty} \frac{r^l}{\sum_{j\in U} a_{+j}} \sum_{j=1}^{N} (a_{+j} - a_{j+}) \big(\frac{a_{ji}}{a_{j+}}\big)^l$$

If $r = 0$, then the PageRank of i depends only on its in-degree. For $r > 0$, page i receives a contribution from another page j via a walk of a given length l: a source page that has a larger in-degree than out-degree would contribute positively, whereas it contributes negatively in the opposite case. A larger r increases the contribution of longer walks.

Betweenness As another type of centrality measure, the shortest-path (SP) *betweenness* of a node i is the fraction of shortest paths, between the pairs of nodes in a graph, which pass through i. In cases of more than one shortest path between a given pair of nodes, each of them is given an equal weight such that the weights sum to one. The denominator of the fraction is $N(N-1)/2$ for undirected graphs.

However, a node may be never on a shortest path, although it may seem intuitively important to the flows in a graph. As can be seen in Figure 6.1, the SP betweenness of ★ is 0, because it is always 'short-circuited' by the two ◯ nodes.

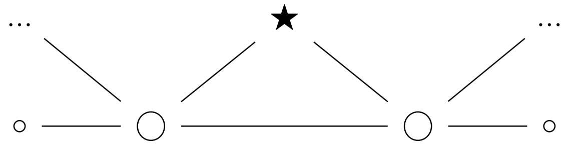

Figure 6.1: An illustration for SP betweenness.

Roughly speaking, the random-walk (RW) betweenness of a node is related to the number of times that a random walk between a pair of nodes passes through it, averaged over all the node pairs in a graph. In the example of Figure 6.1, a random walk through either ◯ has the same chance of moving to ★ or the other ◯. Thus, the RW betweenness of ★ will be much higher than its SP betweenness.

Core-periphery structure One may wish to decompose a connected graph into one or several densely-connected cores along with sparsely-connected peripheral nodes. The nodes in a core are heavily interconnected among themselves, whereas the peripheral nodes are sparsely connected among themselves and are adjacent predominantly to the core nodes. Consequently, a random walk that is located at a peripheral node is much less likely to move to another peripheral node in the next step time. Let the persistent probability of a set of nodes, denoted by M, be given by

$$\xi_M = \sum_{i,j\in M} \pi_i p_{ij} / \sum_{i\in M} \pi_i$$

The idea is that the nodes in M tend to be peripheral if ξ_M is small. One can start with M consisting of a node that has the lowest total in and out-degree and one-by-one add nodes to M, such that each time the increment of ξ_M is the minimum amount possible. The nodes entering later into M are deeper into a core of the graph.

Respondent driven sampling (RDS) A group of individuals of interest may be small compared to the general population and lack any good sampling frame. To obtain a reasonable number of observations from the group of interest, the only practical method may be to follow the links between its members, even though such links are not necessarily exclusive to them. A peculiar feature of RDS is that link-tracing takes place when an in-sample individual (current state) recruits one of its neighbours (adjacent nodes), and participation is usually rewarded financially. Moreover, the out-degree of each node needs to be collected, in order to *model* RDS as a random walk. Taking a random jump amounts to selecting a new initial node, if the last recruit fails to supply any individual.

6.2 Targeted random walk

6.2.1 MH-targeted walk

A walk may be said to be *targeted* if its stationary probability π_i is subject to one's choice beyond that of a pure random walk. This includes particularly the possibility and manner of taking random jumps.

As it can be seen from the flow probability at equilibrium (6.3), one can apply an adjustment to the random-walk transition probabilities based on a_{i+}, in order to achieve the *uniform* stationary probabilities by a walk, where

$$\pi = u$$

Take two adjacent nodes i and j. If $a_{i+} < a_{j+}$, then to increase the fraction of time spent on i, compared to that under the random walk, one can introduce a probability of not moving from i to j, when j is initially selected by the random walk, while always accepting a move from j to i whenever i is selected by the random walk. Clearly, to achieve $\pi_i = \pi_j$ given $a_{i+} < a_{j+}$, the probability of accepting a move from i to j should be a_{i+}/a_{j+}. The same holds also when the walk includes random jumps.

Adding an acceptance-rejection mechanism to Markov chains is called the Metropolis-Hastings (MH) method. Let the right-hand side of (6.4) now be the *proposal probabilities*, which are denoted by q_{ij} so that p_{ij} can still denote the transition probabilities. Let the *acceptance probability* for a move from i to j be

$$\psi_{ij} = \min\left\{\frac{q_{ji}}{q_{ij}}, 1\right\}$$

Let the transition probabilities from i be given by

$$\begin{cases} p_{ij} = q_{ij}\psi_{ij} & \text{if } i \neq j \\ p_{ii} = 1 - \sum_{j \neq i} p_{ij} \end{cases}$$

As explained above, these transition probabilities would result in the uniform stationary probabilities $\pi = u$. Such a walk is referred to as the *uniform walk*.

Table 6.1: Acceptance probabilities ψ_{ij} for $i \neq j$.

Situation		q_{ji}/q_{ij}	ψ_{ij}	Undirected G
$a_{i+} = 0$	$a_{j+} = 0$	$N/N = 1$	1	✓
$a_{i+} = 0$	$a_{ji} = 0$	$N(1-r)/N$	$1-r$	✓
$a_{j+} > 0$	$a_{ji} = 1$	$(1-r) + rN/a_{j+}$	1	–
$a_{i+} > 0$	$a_{ij} = 0$	$N/(1-r)N$	1	✓
$a_{j+} = 0$	$a_{ij} = 1$	$\frac{1/N}{(1-r)/N+r/a_{i+}}$	$\frac{1}{1+r(N/a_{i+}-1)}$	–
$a_{i+} > 0$ $a_{j+} > 0$	$a_{ij} = 0, a_{ji} = 0$	$N(1-r)/(1-r)N$	1	✓
	$a_{ij} = 0, a_{ji} = 1$	$\frac{(1-r)/N+r/a_{j+}}{(1-r)/N}$	1	–
	$a_{ij} = 1, a_{ji} = 0$	$\frac{(1-r)/N}{(1-r)/N+r/a_{i+}}$	$\frac{1-r}{1+r(N/a_{i+}-1)}$	–
	$a_{ij} = a_{ji} = 1$ $a_{i+} < a_{j+}$	$\frac{(1-r)/N+r/a_{j+}}{(1-r)/N+r/a_{i+}}$	$\frac{(1-r)+rN/a_{j+}}{(1-r)+rN/a_{i+}}$	✓
	$a_{ij} = a_{ji} = 1$ $a_{i+} > a_{j+}$	$\frac{(1-r)/N+r/a_{j+}}{(1-r)/N+r/a_{i+}}$	1	✓

Table 6.1 provides the details of the acceptance probabilities ψ_{ij} in different situations. Note that we have $\psi_{ij} = 1$ in the third row because $a_{i+} \leq N$, and $\psi_{ij} < 1$ in the third last row for the same reason. If $N \gg \max(a_{i+}, a_{j+})$ in the second last row, then

$$\psi_{ij} \approx \frac{a_{i+}}{a_{j+}}$$

which is the modification needed to turn a random walk into a uniform walk in a connected graph. Every situation is possible in directed graphs; the last column shows several situations that do not arise in undirected graphs.

To obtain a targeted walk with the specified stationary probabilities π, one can use the general MH acceptance-rejection mechanism given by

$$\psi_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\} \tag{6.5}$$

The uniform walk with $\pi_i \equiv 1/N$ is a special case of such *MH-targeted walks*.

One can set $\pi_i \propto 1 + a_{i+}$, resulting in the *degree+1* walk, which accommodates the nodes with 0 out-degree. For a pair of adjacent nodes with $a_{i+} < a_{j+}$, the acceptance probability is then given by

$$\psi_{ij} = \frac{(1-r)(a_{j+}+1)/N + r(1+1/a_{j+})}{(1-r)(a_{i+}+1)/N + r(1+1/a_{i+})} \approx \frac{1+1/a_{j+}}{1+1/a_{i+}} \quad \text{if} \quad N \gg \max(a_{i+}, a_{j+})$$

One can set $\pi_i \propto \pi(y_i)$, which is a function of the y-value associated with each node, resulting in an adaptive sampling method in valued graphs. For instance, suppose y_i is binary, where one is interested in the nodes with $y_i = 1$ but not those with $y_i = 0$. Setting $\pi(1) = 2$ and $\pi(0) = 1$ means that the former is tuned to have a stationary probability twice that of the latter.

6.2.2 Targeted random walk

The MH-modification of random walk requires a strong OP, where one needs to observe all the successors of all the successors of the current node $X_t = i$, i.e. α_j for all $j \in \alpha_i$, before one can make a move, since q_{ji} is needed in (6.5) in addition to q_{ij}. There is a less demanding scheme, which requires only α_i.

Let there be an imaginary node, denoted by $\star \notin U$, which is connected to all the nodes in an undirected graph G, such that a random jump is accomplished by two successive adjacent moves via $\star$. Let $X_t = i$ at time step t, with degree d_i. Let the probability of moving to $\star$ be $r_i = r/(d_i + r)$ and, having moved to $\star$, one then takes immediately another random move away from it to reach $X_{t+1} = j$, for some $j \in U$. Notice that the probability of taking a random jump r_i is not constant across the nodes. The transition probability from $X_t = i$ to $X_{t+1} = j$ is now given by

$$p_{ij} = \begin{cases} \frac{1}{d_i+r}\left(1 + \frac{r}{N}\right) & \text{if } a_{ij} = 1 \\ \frac{r}{d_i+r} \cdot \frac{1}{N} & \text{if } a_{ij} = 0 \text{ including } i = j \end{cases} \tag{6.6}$$

Since

$$d_j + r = \sum_{i \in \alpha_j} \frac{d_i + r}{d_i + r}\left(1 + \frac{r}{N}\right) + \sum_{i \notin \alpha_j} \frac{d_i + r}{d_i + r} \cdot \frac{r}{N}$$

the resulting stationary probability in undirected graphs is given by

$$\pi_i \propto d_i + r \tag{6.7}$$

We shall refer to (6.6) as the *targeted random walk (TRW)*. It has the same stationary probability as the degree+1 MH-targeted walk if $r = 1$, but without the latter's strong requirement on the OP. It is close to pure random walk, if r is a small positive constant, but without the latter's difficulty in graphs with multiple components. We shall still refer to two successive moves over the imaginary node as a random jump under TRW.

6.2.3 Generalised ratio estimator for 1st-order parameter

Given any constants $y_U = \{y_i : i \in U\}$ associated with the nodes of the graph, Let $\mu = \theta/N$ be a 1st-order graph parameter, where $\theta = \sum_{i \in U} y_i$ and $N = |U|$.

Let a targeted walk have stationary probabilities $\pi_i \propto c_i$, where the values c_i may be unknown for the unobserved nodes. Uniform walk is the special case with $c_i \equiv 1$, where $\pi_i \equiv 1/N$ if N is known. Random walk in an undirected connected graph is the special case with $c_i = d_i$, where c_i is unknown for any unvisited node. TRW is another special case, with $c_i = d_i + r$ in undirected graphs.

A walk may be said to be stationary *draw-by-draw* at equilibrium, now that π is the same for each draw. One can use an extraction of n states,

$$\mathfrak{s}_n = \{X_{t_1}, X_{t_2}, ..., X_{t_n}\} \quad \text{with} \quad t_1 < t_2 < \cdots < t_n$$

which need not to be successive. As a convention to allow for $t_n = T$, the OP of walk is applied to X_T generally. A *generalised ratio estimator* of μ is then given by

$$\hat{\mu} = \Big(\frac{1}{n}\sum_{i \in \mathfrak{s}_n} \frac{y_i}{c_i}\Big) / \Big(\frac{1}{n}\sum_{i \in \mathfrak{s}_n} \frac{1}{c_i}\Big) = \sum_{i \in \mathfrak{s}_n} \frac{y_i}{c_i} / \sum_{i \in \mathfrak{s}_n} \frac{1}{c_i} \tag{6.8}$$

This estimator is *approximately* unbiased for μ given sufficiently large n.

One can reduce the within-walk auto-correlations among the states in $\mathfrak{s}_n$ by extracting time steps that are far apart from each other, in order to treat $\mathfrak{s}_n$ approximately as an IID sample when it comes to variance estimation. An alternative is to administer multiple walks independently. It is then simple to average the multiple estimators and use the between-walk variance as the basis for variance estimation, regardless the within-walk auto-correlations of each walk.

6.3 Sampling strategy under TRW sampling

The OP of walk sampling does not ensure the ancestry knowledge of any motif. Using a targeted walk and the estimator (6.8) is a strategy based on the draw-by-draw stationary probability, *not* the sample inclusion probability. A distinct feature is that the initial node, which serves as a singleton sample $s_0 = \{X_0\}$, does not need to have a known selection probability. However, one can only deal with the 1st-order graph parameters in this way. A different approach can be developed for other finite-order graph parameters under *targeted walk sampling (TWS)*, which is based on the stationary successive sampling probabilities sequence-by-sequence.

To keep focus, we shall use TRW from undirected graphs for illustrations and examples below, although the development is valid for any TWS method. Comments about MH-targeted walk sampling are given only occasionally to enhance the exposition.

6.3.1 Sample graph

The definition of sample graph G_s by (1.2) can accommodate the isolated nodes in G, which can be visited by random jumps but are not incident to A_s even when the walk passes through it. Given the OP of TWS is applied to X_T generally, the seed sample is

$$s = \{X_0, X_1, ..., X_T\}$$

Notice that a node can appear more than once in the seed sample of walk sampling, whereas the seed sample nodes are all distinct under TSBS by definition. Under T-step TRW sampling (TTRWS), we observe $\{i\} \times \alpha_i$ and $\alpha_i \times \{i\}$ for any $i \in s$, such that the reference set is

$$s_{\text{ref}} = s \times U \cup U \times s$$

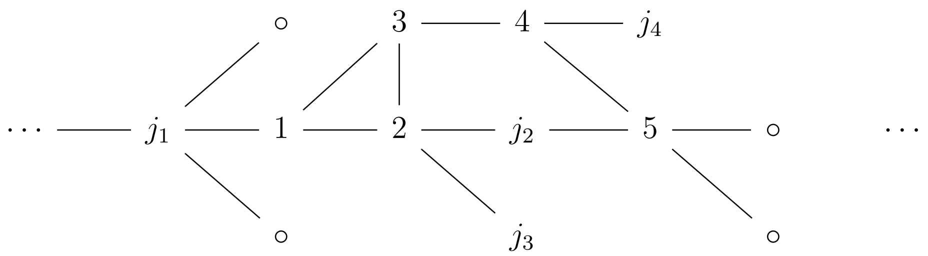

Figure 6.2: An illustration of walk in graph

An illustration of walk in graph is given in Figure 6.2, which passes $\{..., 1, 2, 3, 4, 5, ...\}$ in succession. Some examples of motifs observed by TRW are as below.

- The motif $[\{1, 2\}]$ is observed at $X_t = 1$, since $2 \in \alpha_1$.
- The motif $[\{1, 2, 3\}]$ is observed at $X_{t+1} = 2$ based on $(X_t, X_{t+1}) = (1, 2)$, because we also observe $3 \in \alpha_1$ at $X_t = 1$.
- The motif $[\{1, 2, 3, 4\}]$ is observed at $X_{t+2} = 3$ based on $(X_t, X_{t+1}, X_{t+2}) = (1, 2, 3)$, because we also observe $4 \notin \alpha_1$ at $X_t = 1$ and $4 \notin \alpha_2$ at $X_{t+1} = 2$.

Of course, the nodes M of an observed motif κ do not have to be a sequence in the walk. For example, the motif $[\{1, 3, 5\}]$ is observed at $X_{t+3} = 4$ since $5 \notin \alpha_1 \cup \alpha_3$, assuming the node 5 is unobserved by the walk before that.

An induced motif $[M]$ is observed if $M \times M \subseteq s_{\text{ref}}$, either when all the nodes M belong to the seed sample s or only one node in M falls outside of s.

Let Ω_s be all the observed motifs in G_s. For example, let $X_0 = 1$, $X_4 = 5$ and $T = 4$ in Figure 6.2. We have $s = \{1, 2, 3, 4\}$. Since the nodes in s are all distinct here, the number of observed motifs in $s \times s$ is $2^4 - 1 = 15$. In addition, the nodes $\big(\alpha(s) \cup \beta(s)\big) \setminus s$ are observed, as are the edges in $s \times \big(\alpha(s) \setminus s\big) \cup \big(\beta(s) \setminus s\big) \times s$, such that there are many observed motifs in addition to those in $s \times s$.

MH-targeted walk At each t with $X_t = i$, we observe the edges incident to i as well as j for any $j \in \alpha_i \neq \emptyset$. The reference set of T-step MH-targeted walk sampling is

$$s_{\text{ref}} = s \times U \cup U \times s \quad \text{where} \quad s = \{X_0, \alpha_{X_0}, X_1, \alpha_{X_1}, ..., X_T, \alpha_{X_T}\}$$

For example, let $X_0 = 1$ and $X_T = 4$ and $T = 3$ in Figure 6.2. We have

$$s = \{1, j_1, 2, 3, j_2, j_3, 4, j_4, 5\}$$

after removing the duplicates. The number of observed motifs in $s \times s$ is $2^9 - 1 = 511$.

6.3.2 Stationary successive sampling probability

Let a set of states $M = \{X_{t_1}, ..., X_{t_q}\}$ be given in the order by which the states are sampled, where $t_1 < \cdots < t_q$ and $q = |M|$. At equilibrium, we have

$$\pi_M = \Pr(X_{t_1}, ..., X_{t_q}) = \pi_{X_{t_1}} \prod_{i=1}^{q-1} p(X_{t_i}, X_{t_{i+1}}) \tag{6.9}$$

where the transition probability from X_{t_i} to $X_{t_{i+1}}$ over exactly $t_{i+1} - t_i$ time steps is given by the $(X_{t_i}, X_{t_{i+1}})$-th element of the transition-probability matrix $P^{t_{i+1}-t_i}$, denoted by

$$p(X_{t_i}, X_{t_{i+1}}) = \left[P^{t_{i+1}-t_i}\right]_{X_{t_i} X_{t_{i+1}}}$$

The *stationary sampling probability* (6.9) includes π_i as the 1st-order special case with $M = \{i\}$. Consider two examples. For $(X_t, X_{t+2}, X_{t+4}) = (1, 3, 5)$, we have

$$\pi_{X_t X_{t+2} X_{t+4}} = \pi_1 \big(\sum_{i \in U} p_{1i} p_{i3}\big)\big(\sum_{i \in U} p_{3i} p_{i5}\big)$$

which requires the entire 1st and 3rd rows and 3rd and 5th columns of the matrix P. Whereas for $(X_t, X_{t+1}, X_{t+2}, X_{t+3}) = (1, 2, 3, 4)$, we have

$$\pi_{X_t X_{t+1} X_{t+2} X_{t+3}} = \pi_1 p_{12} p_{23} p_{34}$$

which requires only the transition probabilities of these successive states.

We shall refer to π_M by (6.9) as the *stationary successive sampling probability (S3P)*, when $M = \{X_{t+1}, ..., X_{t+q}\}$ is a *sequence* (of successive states) in a TRW.

Next, in addition to the actual sequences in a TRW, the S3P (6.9) is known up to a proportionality constant for any *hypothetical* sequence $M \subseteq s$, where s is the seed sample of the actual walk. This is because α_i is observed for any $i \in s$, and the sub-matrix of P corresponding to $s \times s$ is known, even though the full matrix P is not available. For instance, given the actual sequence $(X_t, X_{t+1}, X_{t+2}) = (1, 2, 3)$ with $X_T = 5$ under TRW in Figure 6.2, we can also calculate the transition probability $p_{32} p_{21}$ of a hypothetical sequence $(X_t, X_{t+1}, X_{t+2}) = (3, 2, 1)$.

Given T-step walk with seed sample s, let a collection of node sets be

$$\mathcal{C}_s = \{M : M \subseteq s\} \tag{6.10}$$

We shall refer to $\mathcal{C}_s$ by (6.10) as the *generating (sets of) states* of a T-step walk. The subset of $\mathcal{C}_s$ containing parts of the actual walk is given by

$$\mathcal{C}_w = \{\{X_t, ..., X_{t+q}\} : 0 \leq t \leq t+q \leq T\} \tag{6.11}$$

Suppose TRW with $X_0 = 1$ and $X_T = 4$ in Figure 6.2, where $s = \{1, 2, 3, 4\}$. We have

$$\mathcal{C}_w = \big\{\{1\}, \{1,2\}, \{1,2,3\}, \{1,2,3,4\}\big\} \cup \big\{\{2\}, \{2,3\}, \{2,3,4\}\big\} \cup \big\{\{3\}, \{3,4\}\big\} \cup \big\{\{4\}\big\}$$

In addition to the 10 sets in $\mathcal{C}_w$, the S3P is also known up to a proportionality constant for the others in $\mathcal{C}_s$, such as $M = \{4, 2, 1\}$ with $\pi_M = \pi_4 p_{42} p_{21}$ given by (6.9).

6.3.3 Eligible sample motifs

Consider the triangle motif κ with $M = \{1, 2, 3\}$ in Figure 6.2 under TTRWS. The motif is actually observed from $(X_t, X_{t+1}) = (1, 2)$ with S3P $\pi_1 p_{12}$. Let

$$\tau_\kappa = \sum_{(i,j)\in U} \frac{\delta_{i,j}}{\pi_{i,j}} y_{i,j}$$

where $\delta_{i,j} = 1$ if $(X_t, X_{t+1}) = (i, j)$ at equilibrium and 0 otherwise, $\pi_{i,j} = \pi_i p_{ij} = E(\delta_{i,j})$, and $y_{i,j} = 1$ if κ is observed given $(X_t, X_{t+1}) = (i, j)$ and 0 otherwise. We have

$$E(\tau_\kappa) = \sum_{(i,j)\in U} y_{i,j} = 6$$

since $y_{i,j} = 1$ in 6 cases, which are $(X_t, X_{t+1}) =$ (1,2), (2,1), (1,3), (3,1), (2,3) or (3,2). One needs to take into account all these 6 possibilities, when estimating a graph total of triangles, given that one of them is the realised (X_t, X_{t+1}).

Note that a motif can be observed, for which it is infeasible to take such considerations. For example, suppose the triangle motif κ with $M = \{1, 2, 3\}$ is actually observed given $(X_t, X_{t+1}, X_{t+2}) = (1, i, 2)$ for $i \not\in \{1, 2, 3\}$, instead of $(X_t, X_{t+1}) = (1, 2)$. One shall not be able to account for all possible X_{t+1}, as long as $s \neq U$.

Let the sample motif κ in Ω_s be observed from the *actual sampling sequence of states (AS3)* $s_\kappa = (X_t, ..., X_{t+q})$, for some t and $q = |s_\kappa| - 1$. An *equivalent sampling sequence of states (ES3)* of s_κ, denoted by $\tilde{s}_\kappa \sim s_\kappa$, is any possible sequence of states of the same length $|\tilde{s}_\kappa| = |s_\kappa|$, such that the motif κ would be observed given $(X_t, X_{t+1}, ..., X_{t+q}) = \tilde{s}_\kappa$ but not based on any subsequence of $\tilde{s}_\kappa$. In particular, we have $s_\kappa \sim s_\kappa$.

To illustrate with the walk in Figure 6.2, the triangle motif of $M = \{1, 2, 3\}$ has AS3 $(X_t, X_{t+1}) = (1, 2)$, and the other ES3 are $(X_t, X_{t+1}) = (2, 1), (1, 3), (3, 1), (2, 3)$ or $(3, 2)$. The triad motif of $M = \{1, 2, 4\}$ also has $(X_t, X_{t+1}) = (1, 2)$ as its AS3, where the other ES3 are $(X_t, X_{t+1}) = (2, 1), (1, 4), (4, 1), (2, 4)$ or $(4, 2)$.

Take the triad motif of $M = \{2, 4, 5\}$, its AS3 is $(X_t, X_{t+1}, X_{t+2}) = (2, 3, 4)$. Any $(X_t, X_{t+1}, X_{t+2}) = (2, i, 4)$ is its ES3, where $i \not\in \{4, 5\}$. Note that $(2, 4, 4)$ is not an ES3, as the motif is already observed given the subsequence $(X_t, X_{t+1}) = (2, 4)$; similarly for $(2, 5, 4)$. Other ES3s include $(X_t, X_{t+1}, X_{t+2}) = (4, i, 2)$ where $i \not\in \{2, 5\}$, $(2, i, 5)$ where $i \not\in \{4, 5\}$, $(5, i, 2)$ where $i \not\in \{2, 4\}$, $(4, i, 5)$ where $i \not\in \{2, 5\}$, and $(5, i, 4)$ where $i \not\in \{2, 4\}$.

Lemma 6.1. *Under TWS at equilibrium, a motif $\kappa \in \Omega_s$ observed from AS3 s_κ is eligible for estimation, iff all its ES3s belong to the generating states $\mathcal{C}_s$.*

Under TWS at equilibrium, a motif κ whose AS3 is of order greater than 1 can be sampled *sequence-by-sequence*, for which its ES3s constitute the multiplicity. The motif is eligible for estimation if it satisfies the condition of Lemma 6.1, because the S3P of any ES3, $M = \tilde{s}_\kappa$, is known up to a proportionality constant if $\tilde{s}_\kappa \in \mathcal{C}_s$. Thus, estimation of finite-order graph parameters based on TWS sequence-by-sequence generalises estimation of 1st-order parameters based on TWS draw-by-draw.

6.3.4 Generalised ratio estimator

Given any TWS from G, choose a (seed) sample of successive states at equilibrium, which is of size n, after the initial burn-in states, denoted by

$$s = \{X_1, ..., X_n\}$$

Obtain all the observed motifs Ω_s, the generating states $\mathcal{C}_s$ and its subset $\mathcal{C}_w$. For any $\kappa \in \Omega_s$, let s_κ be its AS3, where $s_\kappa \in \mathcal{C}_w$. Denote the set of its ES3s by

$$R_\kappa = \{\tilde{s}_\kappa : \tilde{s}_\kappa \sim s_\kappa\}$$

To illustrate, let $s = \{1, 2, 3, 4\}$ in Figure 6.2. The motif $[\{j_1, 1\}]$ has $s_\kappa = \{1\}$, $R_\kappa = \{1, j_1\}$, such that it is ineligible since $j_1 \not\in s$. The motif of $[\{1, 2, 3\}]$ has AS3 $s_\kappa = \{1, 2\}$ from $(X_1, X_2) = (1, 2)$, as well as $s_\kappa = \{2, 3\}$ from $(X_2, X_3) = (2, 3)$. It can be used twice for estimating a graph total of triangles, with the same

$$R_\kappa = \{(1, 2), (2, 1), (1, 3), (3, 1), (2, 3), (3, 2)\}$$

The motif κ' of $\{1, 2, 4\}$ and κ'' of $\{1, 2, 5\}$ have both the AS3 $(X_1, X_2) = \{1, 2\}$. Since

$$R_{\kappa'} = \{(1, 2), (2, 1), (1, 4), (4, 1), (2, 4), (4, 2)\} \subset \mathcal{C}_s$$

the motif κ' is eligible for estimation, but κ'' is ineligible since $5 \not\in s$, so that

$$R_{\kappa''} = \{(1, 2), (2, 1), (1, 5), (5, 1), (2, 5), (5, 2)\} \not\subset \mathcal{C}_s$$

Let $\theta = \sum_{\kappa \in \Omega} y_\kappa$ be the graph total of interest, where Ω contains all the relevant motifs in the graph. Let an eligible sample motif $\kappa \in \Omega_s$ have AS3 $s_\kappa = (X_t, ..., X_{t+q})$. Let M be a possible sequence of states $(X_t, ..., X_{t+q})$; let $\delta_M = 1$ if M is realised and 0 otherwise, with the associated S3P π_M. Let $I_\kappa(M) = 1$ if $M \in R_\kappa$ and 0 otherwise. Let

$w_{M\kappa}$ be the *incidence weight*, where κ is observed if M is a sample sequence, which is 0 if $I_\kappa(M) = 0$. The corresponding IWE given by

$$\hat{\theta}(X_t, ..., X_{t+q}) = \sum_{\kappa \in \Omega} \sum_{M} \frac{\delta_M}{\pi_M} I_\kappa(M) w_{M\kappa} y_\kappa \tag{6.12}$$

is unbiased for θ if, for any $\kappa \in \Omega$, we have

$$\sum_{M \in R_\kappa} w_{M\kappa} = 1$$

since

$$E\big(\hat{\theta}(X_t, ..., X_{t+q})\big) = \sum_{\kappa \in \Omega} y_\kappa \sum_{M} I_\kappa(M) w_{M\kappa} = \sum_{\kappa \in \Omega} y_\kappa \sum_{M \in R_\kappa} w_{M\kappa}$$

In other words, one can consider R_κ as the ancestors of κ over repeated sampling of $(X_t, ..., X_{t+q})$, just like β_κ of κ under BIGS. While there can be infinitely many possibilities, the two basic choices of $w_{M\kappa}$ are the multiplicity weight given by

$$w_{M\kappa} \equiv 1/|R_\kappa| \tag{6.13}$$

for any M in R_κ, and the *proportional-to-probability weight (PPW)* given by

$$w_{M\kappa} = \pi_M / \pi_{(\kappa)} \qquad \text{and} \qquad \pi_{(\kappa)} = \sum_{M \in R_\kappa} \pi_M \tag{6.14}$$

Given the AS3 s_κ is of the order $|s_\kappa| = q + 1$ and $|s| = n$, there are at most $n - q$ possible estimators, denoted by $\hat{\theta}_t$, for $t = 1, ..., n - q$, based on $(X_t, ..., X_{t+q})$ and given by (6.12). Let $\mathbb{I}_t = 1$ if $(X_t, ..., X_{t+q})$ leads to the observation of at least one motif in Ω, in which case $\hat{\theta}_t$ exists, and 0 otherwise. Provided TWS is stationary sequence-by-sequence, an estimator of θ combining all the $\hat{\theta}_t$ is given by

$$\hat{\theta} = \Big(\sum_{t=1}^{n-q} \mathbb{I}_t \hat{\theta}_t\Big) / \Big(\sum_{t=1}^{n-q} \mathbb{I}_t\Big) \ . \tag{6.15}$$

Insofar as π_M in (6.12) contains an unknown proportionality constant, (6.15) cannot be calculated. However, any function of graph totals that is invariant towards the unknown proportionality constant, when each total is replaced by its estimator (6.15), can be estimated using a generalised ratio estimator, similarly named as (6.8).

For instance, let $\mu = \theta / N_\Omega$, where $N_\Omega = |\Omega|$. A generalised ratio estimator is given by $\hat{\mu} = \hat{\theta} / \hat{N}_\Omega$, where $\hat{N}_\Omega$ is given by (6.12) and (6.15) with $y_\kappa \equiv 1$. This reduces to (6.8) in the special case of $\Omega = U$, $N_\Omega = N = |U|$ and $q = 0$. Or, let

$$\mu = \theta / \theta' \qquad \text{where} \qquad \theta = \sum_{\kappa \in \Omega} y_\kappa \quad \text{and} \quad \theta' = \sum_{\kappa \in \Omega'} y'_\kappa$$

refer to two different kinds of motif in Ω and Ω', respectively. Replacing θ by $\hat{\theta}$ and θ' by

$\hat{\theta}'$, respectively using (6.15), we obtain a generalised ratio estimator $\hat{\mu}$ of μ, as long as $\hat{\mu}$ does not depend on the unknown proportionality constant in the relevant S3Ps.

6.4 Illustrations

Let $G = (U, A)$ be an undirected simple graph with 100 nodes, $N = |U| = 100$. Let $y = 1$ be the value associated with the first 20 nodes $i = 1, ..., 20$, to be referred to as the cases; and let $y = 0$ be the value for the rest 80 nodes, to be referred to as the noncases.

The edges are generated randomly, with different probabilities for a given pair of nodes: (a) if both have $y = 1$, (b) if one of them has $y = 1$ and the other $y = 0$, and (c) if both have $y = 0$. In the resulting graph, there are altogether 299 edges, $|A| = 299$; the cases have an average degree 13.5, and the noncases have an average degree 4.1. The population graph G exhibits a mild core-periphery structure.

This valued graph will be held fixed for the illustrations below.

6.4.1 Convergence to equilibrium

Let $p_{t,i} = \Pr(X_t = i)$, for $i \in U$. We have $p_{t,i} \to \pi_i \propto d_i + r$ by (6.7) under TRW, as $t \to \infty$. How quickly a walk reaches equilibrium is affected by the selection of the initial state X_0. In the extreme case, where $\Pr(X_0 = i) = \pi_i$ for $i \in U$, the walk is at equilibrium from the very beginning. To explore the speed of convergence, consider two other alternatives, where $p_{0,i} = \Pr(X_0 = i) \equiv 1/N$ or $p_{0,1} = 1$. To track the convergence empirically, we use B independent simulations of the TRW to estimate

$$E(Y_t) = \sum_{i \in U} p_{t,i} y_i$$

and compare it to the expectation $E(Y_\infty) = \sum_{i \in U} \pi_i y_i$ at equilibrium.

Table 6.2: $E(Y_t)$ by t, r and initiation of X_0 with 10^5 simulations.

	$r = 1,\ E(Y_\infty) = 0.415$				$r = 0.1,\ E(Y_\infty) = 0.447$			
Initiation	$t = 1$	$t = 4$	$t = 8$	$t = 16$	$t = 1$	$t = 4$	$t = 8$	$t = 16$
$p_{0,i} = \pi_i$	0.420	0.419	0.421	0.414	0.449	0.451	0.450	0.448
$p_{0,i} = 1/N$	0.341	0.394	0.408	0.413	0.355	0.412	0.433	0.449
$p_{0,1} = 1$	0.714	0.481	0.433	0.413	0.741	0.529	0.471	0.448

Table 6.2 shows the results for $t = 1, 4, 8, 16$ and $r = 1, 0.1$, each based on $B = 10^5$ simulations. The equilibrium expectation $E(Y_\infty)$ varies with r, which affects the transition probabilities. Of the two choices here, the value $r = 1$ yields the degree+1 walk, whereas the value $r = 0.1$ tunes the walk closer to pure random walk.

It can be seen that TRW stays at equilibrium if $p_{0,i} = \pi_i$. For a given value of r, the differences as t varies provide a tangible appreciation of the simulation error, when each result is obtained from 10^5 simulations. Under the current set-up, convergence to equilibrium is apparently achieved at $t = 16$ already, whether the initial X_0 is selected completely randomly from U given $p_{0,i} = 1/N$, or fixed at $i = 1$ given $p_{0,1} = 1$. Neither does the speed of convergence vary much for the values of r here.

6.4.2 Estimation of case prevalence

Let $\mu = \sum_{i \in U} y_i/N$ be the population case prevalence. Let $s = \{X_0, ..., X_T\}$ be the states obtained by TTRWS, where X_0 is drawn with $p_{0,i} = 1/N$. Apply (6.8) to s yields $\hat{\mu}$. The burn-in stage is quite short in the present setting. In any case, using all the states, without removing the burn-in stage before a walk reaches equilibrium, is instructive for appreciating the convergence of the generalised ratio estimator $\hat{\mu}$.

For given T and r, generate TRW independently B times, each resulting in a replicate of the estimator $\hat{\mu}$. The mean of the B replicates is an estimate of $E(\hat{\mu})$ under TTRWS, and the variance of them is an estimate of $V(\hat{\mu})$. In addition, calculate the naïve variance estimate $s_T^2/(T + 1)$, where $s_T^2 = \sum_{t=0}^{T}(y_{X_t} - \bar{y})^2/T$ and $\bar{y}$ are the sample variance and mean of $\{y_{X_t} : t = 0, ..., T\}$, respectively. This is not a consistent variance estimator due to the within-walk auto-correlations. Moreover, let $\mu(1 - \mu)/(T + 1)$ be the theoretical variance when one can estimate μ based on an IID sample, which is drawn randomly and with replacement from $\{y_i : i \in U\}$.

In particular, denote by ψ the *traverse* of the walk, given as the ratio between the number of distinct nodes visited by the walk and $N = |U|$, which indicates how extensively the walk has travelled through the population graph.

Table 6.3: Estimation of case prevalence $\mu = 0.2$ under TTRWS, 10^3 simulations.

	T	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	Naïve SD	SD-IID	ψ
$r = 1$	50	0.200	0.081	0.067	0.056	0.346
	100	0.199	0.059	0.048	0.040	0.538
	500	0.200	0.027	0.022	0.018	0.938
	1000	0.199	0.019	0.016	0.013	0.987
$r = 0.1$	50	0.204	0.091	0.067	0.056	0.321
	100	0.205	0.068	0.049	0.040	0.501
	500	0.201	0.031	0.022	0.018	0.893
	1000	0.201	0.022	0.016	0.013	0.959

Table 6.3 gives the results for $T = 50, 100, 500, 1000$, $r = 1$ or 0.1, each based on $B = 1000$ simulations of the TTRWS. The mean of the B naïvie SD (square root of variance) estimates is given, which clearly underestimates the true variance of $\hat{\mu}$. Notice that in the current setting, as the length of walk T increases, the absolute bias of naïve variance estimation is reduced but not the relative bias. Without applying any adaptive observation procedure, TRW sampling entails a loss of efficiency compared to standard SRS, as can be seen from SD-IID based on an IID sample.

Note that in the population graph here, the case nodes have larger degrees than the noncase nodes, such that the stationary probability π_i is not independent of y_i, and TRW sampling is informative in this sense. Nevertheless, informative sampling as such is not an issue for design-based estimation of graph parameters.

The consistency of $\hat{\mu}$ under TRW sampling is already evident at $T = 50$, even without removing the initial burn-in states. The last column shows the average of the traverse ψ over the B simulations. In the current setting, a TRW of length $T = 50$ is expected to visit only about a third of the 100 nodes in the population graph, whereas even a walk of length $T = 1000$ can not always reach all the nodes.

How quickly TRW traverses the population graph depends on the isolated nodes that can only be visited by random jumps, which are relatively infrequent. The probabilities of random jumps are reduced given small r. The TRW moves somewhat more slowly through the population graph, given $r = 0.1$ than $r = 1$, as is the convergence of $\hat{\mu}$, which can be seen by comparing the corresponding SD($\hat{\mu}$) given $r = 0.1$ or $r = 1$.

6.4.3 Estimation of a 3rd-order graph parameter

Let $\mu = \theta/\theta'$, where θ is the total number of triangles among cases with all the three nodes having $y = 1$, and θ' is that of the other triangles with at least one noncase node. The larger the value of μ, the higher is the transitivity among cases compared to the overall transitivity in the graph. We have $\mu = 4.667$ in this population graph.

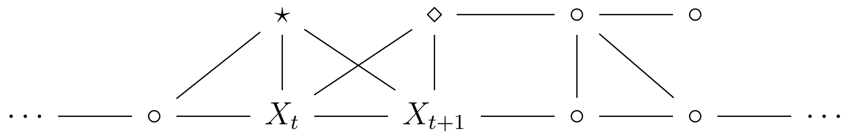

Figure 6.3: Illustration for triangle motifs under TRW.

Given $(X_t, X_{t+1}) = (i, j)$ of two adjacent nodes under TRW, one observes all the triangles that include i and j as two of the nodes. As illustrated in Figure 6.3, both the $\star$-triangle $[\star ij]$ and $\diamond$-triangle $[\diamond ij]$ are observed from the AS3 $(X_t, X_{t+1}) = (i, j)$. For the summations of (6.12), we have $\delta_M = 1$ iff $M = (i, j)$ and $I_\kappa(M) = 1$ if κ is either of these triangles, such that both contribute to $\hat{\theta}_t$ in (6.15) if they both belong to Ω.

Given $(X_t, X_{t+1}) = (i, j)$ as the AS3, the ES3s of any sampled triangle are the six possible adjacent moves along the triangle. Under TRW (6.6) and (6.7), the S3P is

$$\pi_i p_{ij} \equiv 1 + \frac{r}{N}$$

such that (6.13) and (6.14) lead to the same incidence weight $w_{M\kappa} \equiv 1/6$ here.

Table 6.4: Estimation of $\mu = \theta/\theta' = 4.667$ under TTRWS, 1000 simulations.

$r = 0.1$				$r = 6$			
T	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ	T	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ
100	6.119 (0.142)	4.498	0.498	100	6.362 (0.160)	5.069	0.606
500	4.737 (0.028)	0.893	0.893	500	4.805 (0.034)	1.075	0.983
1000	4.669 (0.019)	0.593	0.958	1000	4.704 (0.022)	0.702	0.999

Table 6.3 gives the results for $T = 100, 500, 1000$, $r = 0.1$ or 6, each based on $B = 1000$ simulations of TTRWS. The initial X_0 is selected with $p_{0,i} = \pi_i$ to avoid any details of handling the burn-in states. There is a small probability that a particular run of TRW does not yield any triangle when $T = 50$, making it impossible to compute $\hat{\mu}$. Any choice of $T < 100$ is omitted for this reason.

The results given $r = 0.1$ are in the left part of Table 6.4 and those given $r = 6$ to the right. Since the average degree is about 6 in the population graph here, setting $r = 6$ in (6.6) makes a random jump on average at least as probable as an adjacent move at each time step. This raises the traverse of the walk, e.g. TRW of length $T = 1000$ can now be expected to cover almost the whole population graph.

The convergence of the estimator $\hat{\mu}$ seems not greatly affected by r, although intuitively a large value of r is unlikely to be efficient, as the order of graph parameter increases. Given either r, convergence is at least almost the case given $T = 1000$, where each value in the parentheses in Table 6.4 is the estimated simulation error of Mean$(\hat{\mu})$. Clearly, the walk needs to be longer for estimating this 3rd-order graph parameter than for the population case prevalence. This is not surprising because not every two successive states (X_t, X_{t+1}) correspond to an adjacent move, nor does one necessarily observe any triangle based on every adjacent move. In contrast, every state X_t contributes to the estimation of a 1st-order graph parameter.

Bibliographic notes

Masuda et al. (2017) provide a comprehensive review of random walks and diffusion on networks, albeit without any particular focus on the sampling theory. Boyd et al. (2004) provide some results concerning the rate of convergence for pure random walks in a graph. No general results seem to be available for targeted walks, where the stationary probability π_i is not exclusively determined by the degree of the node in undirected graphs, and the walk allows for random jumps in addition.

Salganik and Heckathorn (2004) propose RDS for hard-to-catch populations, inspired by random walk in graphs. However, it is mostly unrealistic to assume that respondent-driven observation can satisfy the requirement of the random-walk OP. In applications of RDS, therefore, the estimation and inference need to rely on model assumptions rather than the theoretical stationary probabilities.

Until recently, the sampling theory for random walk in graphs has only dealt with the estimation of 1st-order graph parameters, such as the population case prevalence. In particular, Thompson (2006a) consider the generalised ratio estimator under random walk with jumps and targeted MH walks. The more practical algorithm of TRW is given by Avrachenkov et al. (2010).

Thompson (2006b) develops adaptive web sampling for the estimation of 1st-order graph parameters. Web sampling is a kind of hybrid of breadth-first SBS and depth-first TWS. However, the sampling strategy of Thompson (2006b) does not scale to large graphs and samples. This is an interesting topic for future research.

The BIGS-IWE strategy can be adapted to TWS sequence-by-sequence. For the 3rd-order parameter by TTRWS illustrated above, the two basic choices of incidence weight coincide. But this is not the general case. For instance, let $[\{i, j, g, h\}]$ be a 4-cycle, which can be observed based on AS3 consisting of two successive adjacent moves, say, $(X_t, X_{t+1}, X_{t+2}) = (i, j, g)$. The corresponding S3P is given by

$$\pi_i p_{ij} p_{jg} = \frac{1}{d_j + r}\big(1 + \frac{r}{N}\big)^2$$

which varies with the degree of the second node in the sequence, and is no longer a constant over all the ES3s. Thus, sensible choice of the BIGS-IWE strategy under TWS is a topic that can be explored further.

As another intriguing observation, Table 6.4 shows that the traverse of TRW is often 1 given $T = 1000$ and $r = 6$, in which case the whole population graph is observed. Obviously, had one knew that the traverse is 1, one could have computed μ exactly without any error; but since the traverse is unknown when $N = |U|$ is unknown, one cannot readily take advantage of the situation. A topic for future research is how to estimate the traverse and then adjust the estimator accordingly.

Bibliography

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

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

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

[4] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18: 105-110.

[5] Boyd, S., Diaconis, P. and Xiao, L. (2004). Fastest mixing Markov chain on a Graph. *SIAM Review*, 46:667-689.

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

[7] Chung, F.R.K. (1997). *Spectral Graph Theory.* Providence, RI: American Mathematical Society.

[8] Cochran, W.G. (1977). *Sampling Techniques (3rd ed.).* New York: Wiley.

[9] Dryver, A.L. and Thompson, S.K. (2005). Improved unbiased estimators in adaptive cluster sampling. *Journal of the Royal Statistical Society, Series B*, 67:157-166.

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

[11] Frank, O. (1977a). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4:81-89.

[12] Frank, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, 4:178-180.

[13] Frank, O. (1977c) Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1:235-264.

[14] 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.

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

[16] Frank, O. (1980a). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4:45-50, 1980.

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

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

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

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

[21] Godambe, V.P. and Joshi, V.M. Admissibility and Bayes Estimation in Sampling Finite Populations. *The Annals of Mathematical Statistics*, 36:1707-1722.

[22] Goldenberg, A., Zheng, A.X., Fienberg, S.E. and Airoldi, E.M. (2010). A Survey of Statistical Network Models. *Foundations and Trends® in Machine Learning*, 2:129-233.

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

[24] Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663-685.

[25] Kaiser, L. (1983). Unbiased estimation in line-intercept sampling. *Biometrics*, 39:965-976.

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

[27] Lavalleè, P. (2007). *Indirect Sampling*. Springer.

[28] 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`

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

[30] 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.

[31] Patone, M. (2020) *Topics of Statistical Analysis with Social Media Data*. Unpublished PhD Thesis.

[32] Patone, M. and Zhang, L.-C. (2020) Incidence weighting estimation under bipartite incidence graph sampling. `arXiv:2004.04257v1`

[33] Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37:81-91.

[34] Salganik, M.J. and Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193-239.

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

[36] Sirken, M.G. (2004). Network sample survey of rare and elusive populations: a historical review. In *Preceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to Reach Population.*

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

[38] Sirken, M.G. and Levy, P.S. (1974). Multiplicity estimation of proportion based on ratios of random variable. *Journal of the American Statistical Association*, 69:68-73.

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

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

[41] Thompson, S.K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometics*, 47:1103-1115.

[42] Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32:11-24.

[43] Thompson, S.K. (2006b). Adaptive Web Sampling. *Biometrics*, 62:1224-1234.

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

[45] Vincent, K. and Thompson, S.K. (2017). Estimating population size with link-tracing sampling. *Journal of the American Statistical Association*, 112:1286-1295.

[46] Zhang, L.-C. (2021). Graph sampling: An introduction. *The Survey Statistician*, 83:27-37.

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

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

[49] Zhang, L.-C. and Patone, M. (2017) Graph sampling. *Metron*, 75:277.