Chapter 5
Sampling and Estimation in Network Graphs

Frequently it is the case that relational information is observed on only a portion of a complex system being studied, and the network resulting from such measurements may be thought of as a sample from a larger underlying network. If the goal is to use the sampled network data to infer properties of the underling network, this task may be approached using principles of statistical sampling theory. However, sampling in a network context introduces various potential complications. In this chapter we formalize the problem of sampling and estimation in network graphs, describe a handful of common network sampling designs, and develop estimators of a number of quantities of interest.

5.1 Introduction

The material in the previous two chapters has been developed under the implicit assumption that the network graph being mapped and explored is itself the primary object of interest. However, as was mentioned in Section 3.2, this perspective is one of 'measurement as enumeration' and, in particular, ignores the possible effects of sampling.

Although there are many cases in which a network graph is essentially observed in its entirety, arguably sampled graphs may be more the rule than the exception in practice. For example, in social networks, while it may be possible to fully construct the friendship network among school children in a small classroom, it may be cumbersome to attempt to do so for all employees in a large corporation. Similarly, as we saw in the case study on mapping the Internet, in Section 3.5.2, sampling is a necessary part of Internet topology studies.

Formally, assume there is a system under study that may be represented by a network graph G, which we will call the *population graph.* Suppose in addition, however, that instead of having all of G available to us, we take measurements that effectively yield a sample of vertices and edges, which we compile into a graph $G^* = (V^*, E^*)$. We will refer to G^* as a *sampled graph.* The sampled graph G^* will

E.D. Kolaczyk, *Statistical Analysis of Network Data*, Springer Series in Statistics,
DOI 10.1007/978-0-387-88146-1_5, © Springer Science+Business Media, LLC 2009

often be a subgraph of G (i.e., $V^* \subseteq V$ and $E^* \subseteq E$), although this will not always be the case. For example, in situations where there is error in assessing the existence of vertices or edges, there is chance of spurious observations.

Now suppose that there is a particular characteristic of G, denoted $\eta(G)$, that is of interest. For example, $\eta(G)$ might be a structural characteristic of G, such as the number of edges N_e, the average degree, or the distribution of vertex betweenness centrality scores. Similarly, it might be a summary of some quantity with which the nodes or edges in G are decorated, such as the proportion of men with more female than male friends in a social network. If G is sampled, then typically it will be impossible to recover the exact value of $\eta(G)$ from only the partial information supplied by G^*. The question thus arises as to whether we may still obtain a useful estimate of $\eta(G)$, say $\hat{\eta}$, from G^*.

Intuitively, it is attractive to think that we might simply estimate $\eta(G)$ by $\hat{\eta} = \eta(G^*)$. That is, we might use a 'plug-in' method and estimate the characteristic of interest by the value of that characteristic observed in the sampled graph G^*. This approach is in fact implicitly what is used in any network study that asserts that properties of an observed network graph are indicative of those same properties for the graph of the network from which the data were sampled. And certainly many familiar estimators in general practice are plug-in estimators of this type. Sample means, standard deviations, and quantiles, for example, are all both natural and valid estimates of their population equivalents under standard assumptions of a sample with independent and identically distributed observations.

Unfortunately, in estimating graph characteristics from sample graphs, this line of reasoning can often go awry, as the following example illustrates.

Example 5.1 (Estimating Average Degree). Suppose that the characteristic of interest is the average degree of a graph G,

$$\eta(G) = (1/N_v) \sum_{i \in V} d_i \quad . \tag{5.1}$$

Let our sample graph G^* be based on the n vertices $V^* = \{i_1, \ldots, i_n\}$, and denote its observed degree sequence by $\{d_i^*\}_{i \in V^*}$. The plug-in estimator of $\eta(G)$ in (5.1) is just the average of the observed degree sequence,

$$\hat{\eta} = \eta(G^*) = (1/n) \sum_{i \in V^*} d_i^* \quad . \tag{5.2}$$

To evaluate this estimator, we consider two sampling designs by which G^* might be obtained. In both cases, we begin with a simple random sample without replacement[1] of n vertices $V^* = \{i_1, \ldots, i_n\}$. Then, in Design 1, for each vertex $i \in V^*$, we observe all edges $\{i, j\} \in E$ involving i; each such edge becomes an element of E^*. On the other hand, in Design 2, we only observe, for each pair $i, j \in V^*$, whether or not $\{i, j\} \in E$; if it is, that edge becomes an element of E^*. Hence, both designs

[1] That is, elements are drawn sequentially and uniformly at random from a population in such a way as to avoid sampling any element more than once. See Example 5.2.

consist of two steps – first sampling a set of vertices V^* and then observing a set of edges E^* – but differ in the manner in which the edges are observed. In either case, the final sampled graph is simply $G^* = (V^*, E^*)$.

Figure 5.1 shows histograms showing the results of calculating the sample average $\hat{\eta} = \eta(G^*)$ under each of these two sampling designs, in the case where the true graph G is taken to be the network of protein interactions in yeast introduced in Section 4.2.1.1. Recall that G consists of $N_v = 5,151$ vertices and $N_e = 31,201$ edges; it has an average degree of $\eta(G) = 12.115$. A random sample of $n = 1,500$ vertices was drawn, and edges were sampled per the specifications of Designs 1 and 2, respectively. This process was repeated for $10,000$ trials.

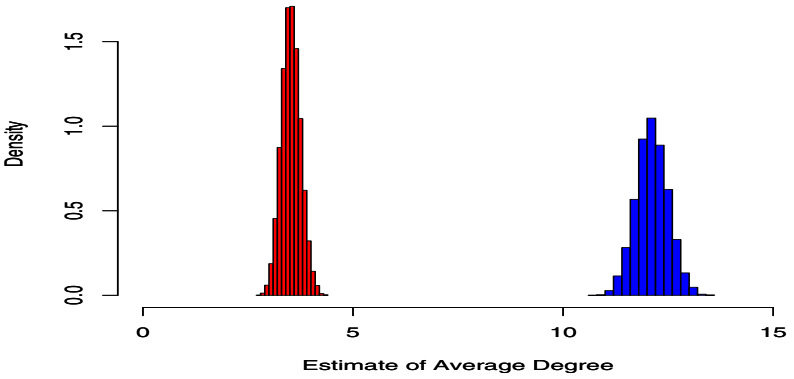

Fig. 5.1 Histograms of estimated average degree in the yeast protein interaction network, based on sampling under Design 1 (blue) and Design 2 (red), over $10,000$ trials.

Looking at the figure, it is easy to see that under Design 1 the plug-in estimator $\hat{\eta}$ is quite accurate, with a mean of 12.117 and a standard error of 0.3797. However, under Design 2 it has a substantial bias, with a mean of only 3.528, although its standard deviation is a bit smaller, at 0.2260. The difference in the performance of the estimators in this example is, of course, a function of the difference in the observed degrees $\{d_i^*\}_{i \in V^*}$, induced by the difference in which edges are observed. In the case of Design 1, we observe the actual degree of a vertex $i \in V^*$ i.e., $d_i^* = d_i$. But in the case of Design 2, we observe a degree that typically under-shoots the actual degree, by a factor of roughly n/N_v i.e., $d_i^* \approx n d_i / N_v$. Hence, despite the fact that in both cases vertices are drawn through simple random sampling, in the latter design our estimate of the average degree is an under-estimate. □

While the sampling designs in Example 5.1 are relatively simple, they are by no means uncommon, and rather than being an exceptional case, the behavior witnessed above has been seen to occur under a variety of sampling designs for various

choices of characteristic $\eta(G)$ (e.g., Lee, Kim, and Jeong [253]). In the specific context of protein interaction networks, Han et al. [192] have documented analogous results using sampling designs more representative of experimental protocols in proteomics.

We shall see in this chapter that, when viewed from the perspective of classical statistical sampling theory, such problems with network graph estimation come as little surprise and, indeed, can be compensated for in many cases through appropriately designed estimation strategies. In Example 5.1, for instance, adjusting the plug-in estimator $\hat{\eta}$ upward by a factor of $N_v/n = 5,151/1,500 \approx 3.434$ yields an estimator with mean 12.115. In this case the correction is easy, but in general, deriving such correction strategies can be more challenging, depending on the manner in which the topology of the graph G, the characteristics of $\eta(\cdot)$, and the nature of the sampling design interact.

5.2 Background on Statistical Sampling Theory

Statistical inference in classical sampling theory is based upon a slightly different paradigm than, say, the likelihood-based or Bayesian frameworks described in Chapter 2. In particular, the standard setup assumes that observations are made without measurement error, and hence that the only source of randomness is the sampling design. Inferential procedures constructed under these assumptions are referred to as design-based procedures. In this chapter we adopt the design-based perspective throughout, beginning here with a brief review of some basic concepts and techniques, which will provide us with an appropriate set of tools to attack the network graph estimation problem. In order to incorporate measurement errors (and other such sources of randomness) into an analysis, it is generally necessary to specify a model, leading to so-called model-based procedures. We will revisit and illustrate the distinction between design-based and model-based procedures, in the context of network graph estimation, in Section 6.2.4.1.

5.2.1 Horvitz-Thompson Estimation for Totals

Suppose we have a population $\mathscr{U} = \{1, \ldots, N_u\}$ of N_u units (e.g., people, animals, widgets, etc.), and suppose that with each unit $i \in \mathscr{U}$ there is associated a value y_i of interest (e.g., height, age, gender, etc.). Let $\tau = \sum_i y_i$ and $\quad = \tau/N_u$ be the total and average values of the y's in the population. Finally, let $S = \{i_1, \ldots, i_n\}$ be a sample of n units from $\mathscr{U}$, and assume that we observe y_i for each $i \in S$.

Estimation of a population total τ or average $\quad$ from a random sample is an ubiquitous task. In the canonical case in which S is chosen by drawing n units uniformly from $\mathscr{U}$, with replacement, a natural estimate of $\quad$ is simply the sample mean $\bar{y} = (1/n)\sum_{i \in S} y_i$. The corresponding estimate of the total is then just

$\hat{\tau} = N_u \bar{y}$. These estimates are unbiased (i.e., $\mathbb{E}(\bar{y}) =$ and $\mathbb{E}(\hat{\tau}) = \tau$), in the sense that the average of their values over all possible samples of size n are the parameters they are estimating. The variances of these estimators take the forms $\mathbb{V}(\bar{y}) = \sigma^2/n$ and $\mathbb{V}(\hat{\tau}) = N_u^2\sigma^2/n$, where σ^2 is the variance of the values y in the full population $\mathscr{U}$. These variances can themselves be estimated in an unbiased fashion through the quantities $\widehat{\mathbb{V}}(\bar{y}) = s^2/n$ and $\widehat{\mathbb{V}}(\hat{\tau}) = N_u^2 s^2/n$, respectively, where $s^2 = (1/(n-1))\sum_{i\in S}(y_i - \bar{y})^2$ is the sample variance.

In many situations, however, the sample S is far from being a simple random sample with replacement. In particular, often some units are more likely than others to be included in a sample, either by design or by accident. A marketing survey might favor calling larger households. A census survey may have more trouble locating homeless people than homeowners. As we will see below, similar scenarios can often arise in the sampling of networks. Sampling under designs of this sort is called *unequal probability sampling*.

When sampling with unequal probabilities, the sample mean can be a poor choice of estimator, in that it will be a biased estimator. The *Horvitz-Thompson estimator* is constructed in a manner that remedies this problem, through the use of weighted-averaging. Suppose that, under a given sampling design, each unit $i \in \mathscr{U}$ has probability π_i of being included in a sample of size n. Let $S \subset \mathscr{U}$ now be the set of distinct units in the sample. Then the Horvitz-Thompson estimate of the total τ takes the form

$$\hat{\tau}_\pi = \sum_{i\in S} \frac{y_i}{\pi_i} \quad . \tag{5.3}$$

The corresponding estimate of the mean is obtained as $\hat{}_\pi = (1/N_u)\hat{\tau}_\pi$.

The estimator $\hat{\tau}_\pi$ is an unbiased estimate of τ, assuming $\pi_i > 0$ for all $i \in \mathscr{U}$. To see this, let $Z_1, \ldots, Z_{N_u}$ be a set of binary random variables such that $Z_i = 1$ if unit i is in S, and zero otherwise. Then, since

$$\mathbb{E}(\hat{\tau}_\pi) = \mathbb{E}\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right) = \mathbb{E}\left(\sum_{i\in\mathscr{U}}\frac{y_i}{\pi_i}Z_i\right) = \sum_{i\in\mathscr{U}}\frac{y_i}{\pi_i}\mathbb{E}(Z_i) \quad , \tag{5.4}$$

and $\mathbb{E}(Z_i) = \mathbb{P}(Z_i = 1) = \pi_i$, by definition, the result follows. Similarly, $\hat{}_\pi$ is an unbiased estimate of .

If we define π_{ij} to be the probability that units i and j are both in the sample S, with $\pi_{ij} = \pi_i$ for convenience when $i = j$, then the variance of our estimator $\hat{\tau}_\pi$ can be expressed as

$$\mathbb{V}(\hat{\tau}_\pi) = \sum_{i\in\mathscr{U}}\sum_{j\in\mathscr{U}} y_i y_j \left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) \quad , \tag{5.5}$$

and may itself be estimated in an unbiased fashion by the quantity

$$\widehat{\mathbb{V}}(\hat{\tau}_\pi) = \sum_{i\in S}\sum_{j\in S} y_i y_j \left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right) \quad , \tag{5.6}$$

assuming $\pi_{ij} > 0$ for all pairs i, j. Analogous quantities for the estimator $\hat{}_\pi$ follow from the expressions $\mathbb{V}(\hat{}_\pi) = (1/N^2)\mathbb{V}(\hat{\tau}_\pi)$ and $\widehat{\mathbb{V}}(\hat{}_\pi) = (1/N^2)\widehat{\mathbb{V}}(\hat{\tau}_\pi)$. Note that in the case that inclusion of units i and j are independent events, we have $\pi_{ij} = \pi_i \pi_j$, and some simplication of the expressions in (5.5) and (5.6) is possible.

Example 5.2 (Simple Random Sampling Without Replacement). Consider the case where n units are sampled from $\mathscr{U}$ uniformly, but without replacement. That is, i_1 is chosen uniformly from $\mathscr{U}$ and then set aside. Next, i_2 is chosen uniformly from $\mathscr{U} \setminus \{i_1\}$, and so on, until finally i_n is chosen uniformly from $\mathscr{U} \setminus \{i_1, \ldots, i_{n-1}\}$.

As there are $\binom{N_u}{n}$ possible samples of size n that may be chosen in this manner, and $\binom{N_u-1}{n-1}$ samples that may be chosen to include a given unit i, it follows that

$$\pi_i = \frac{\binom{N_u-1}{n-1}}{\binom{N_u}{n}} = \frac{n}{N_u} \quad , \tag{5.7}$$

for all $i \in \mathscr{U}$. Arguing similarly, we have $\pi_{ij} = [n(n-1)]/[N_u(N_u-1)]$.

Therefore, it follows that under this design the Horvitz-Thompson estimates of the total and mean have the form

$$\hat{\tau}_\pi = N_u \bar{y} \quad \text{and} \quad \hat{}_\pi = \bar{y} \ . \tag{5.8}$$

Note that these are the same unbiased estimators described at the start of this section for simple random sampling *with* replacement. However, the variance of, for example, $\hat{\tau}_\pi$ may be shown to be $N_u(N_u - n)\sigma^2/n$, which is strictly less than the value $N_u^2\sigma^2/n$ we saw earlier under simple random sampling with replacement.

In other words, under simple random sampling, the sample mean $\bar{y}$ is unbiased whether sampling with and without replacement. However, $\bar{y}$ has a lower variance when sampling without replacement, as the same number of samples n is used more efficiently by enforcing that all units in S be distinct. □

Example 5.3 (Probability Proportional to Size Sampling). In many contexts, while sampling is done with replacement, the units are not sampled with equal probability at each stage. Rather, for each of the n times a unit is selected from $\mathscr{U}$, the unit is selected with respect to a probability distribution $\{p_1, \ldots, p_{N_u}\}$ on $\mathscr{U}$.

This situation naturally arises when units are selected from the population with probabilities p_i directly proportional to the values c_i of some characteristic. For example, households might be selected for a marketing survey by drawing names from a database, in which case those households with more members in the database have a larger chance of being included. Alternatively, a study of wildlife in a remote geographical region might be conducted by flying an aircraft over adjacent strips of land of a given width. But natural geographic features often result in strips being of different lengths, and hence different areas, which is likely to make the chance of seeing a given type of animal vary. Sampling conducted in this manner is called *probability proportional to size* (PPS) sampling.

Because sampling is done with replacement, it is easy to see that the inclusion probabilities take the form $\pi_i = 1 - (1 - p_i)^n$, where $p_i = c_i / \sum_i c_i$, and hence are unequal if the p_i are unequal. The Horvitz-Thompson estimators of τ and are therefore more appropriate here than the sample mean, particularly if the p_i are decidedly non-uniform. Accompanying estimates of the variance of these estimators can be computed using the values $\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$. □

As can be seen, Horvitz-Thompson estimators are in principle quite general in their applicability. In reality, of course, they are limited to those contexts in which the values of the π_i may be computed for the underlying sampling design. Such computations tend to be noticeably more difficult for designs that sample without replacement. Nevertheless, we will see in Section 5.4 a number of different network graph sampling designs for which the values π_i may be computed. And even in cases for which it appears difficult or impossible to compute these inclusion probabilities, the Horvitz-Thompson framework nonetheless provides a perspective by which valuable insight may be gained.

5.2.2 Estimation of Group Size

A special type of population total that is often of interest is the size of a group. Note, for example, that in the previous material it was implicitly assumed that N_u, the size of the population $\mathscr{U}$, is known. Many times this is simply not true. In fact, in some cases estimating N_u is an important goal in and of itself. For example, there are many populations that are 'hard to find' and yet whose size is important to know for planning purposes, such as populations of endangered animal species or populations of humans at risk for a particular rare disease or condition.

In principle, we may write $N_u = \sum_{i \in \mathscr{U}} 1$, which would suggest the Horvitz-Thompson estimator $\hat{N}_u = \sum_{i \in S} \pi_i^{-1}$. Unfortunately, as we have seen in Examples 5.2 and 5.3, knowledge of N_u is typically needed to compute the π_i, which makes this approach infeasible.

Instead, alternative techniques have been developed for this particular estimation problem, the primary example of which is the class of *capture-recapture* estimators. The simplest version of capture-recapture involves two stages of simple random sampling without replacement, yielding two samples, say S_1 and S_2. In the first stage, the sample S_1 of size n_1 is taken, and all of the units in S_1 are 'marked.' Marking might correspond to literally tagging a fish or animal, or simply noting the ID number of a record in a database. All of the units in S_1 are then 'returned' to the population – either literally or figuratively – and, at the second stage of sampling, a sample of size n_2 is taken from $\mathscr{U}$. The value

$$\hat{N}_u^{(c/r)} = \frac{n_2}{m} n_1 \tag{5.9}$$

is then used as an estimate of N_u, where $m = |S_1 \cap S_2|$ is the number of marked units observed in the second sample. An estimator of the variance is

$$\widehat{\mathbb{V}}(\hat{N}_u^{(c/r)}) = \frac{n_1 n_2 (n_1 - m)(n_2 - m)}{m^3} \quad . \tag{5.10}$$

The capture-recapture estimator in (5.9) may be seen to adjust upward the size n_1 of the first sample by one over the factor m/n_2, an indication of what fraction of the overall population was marked. It has been derived using a number of different arguments. For example, if n_1 and n_2 are fixed in advance, m has a hypergeometric distribution, and the integer part of $\hat{N}_u^{(c/r)}$ can be shown to be the corresponding maximum likelihood estimate of N_u. Similar estimators have been derived under various other models and assumptions, allowing for such changes as random n_1, n_2, unequal probability sampling of units, changes in inclusion probabilities between the first and second stages of sampling, and more than two stages of sampling.

In Section 5.5 we will consider the problem of estimating the size of a 'hard to find' sub-population, not using capture-recapture sampling but rather a certain type of network sampling. Nevertheless, we will see that useful estimators may be derived with a form quite similar to that in (5.9) .

An important related problem of group-size estimation is the so-called *species problem*. This problem takes its name from biology, where the standard caricature is that of a lonely biologist sitting in the middle of a forest observing animals that pass within view. While the biologist records each animal sighted, in the end it is actually the number of species of animals in the forest, say N_s, that is of interest. Despite the biological motivation of its name, this problem arises quite widely. For example, the estimation of the size of an author's apparent vocabulary from published works (e.g., "How many words did Shakespeare know?"), or the estimation of the number of ancient coins minted by a society based on archaeological finds, have both been cast as species problems.

In the special case that each species has the same proportion of members in $\mathscr{U}$, under random sampling the problem of estimating N_s arguably is not much different from that of estimating N_u. Of course, the assumption of equal proportions of members among species is typically not true. In reality, unfortunately, unless there is prior knowledge to the contrary, there is always the possibility that there are an arbitrary number of species in the population in arbitrarily small proportions (given a population of sufficient size). This fact allows for the species problem to potentially be quite ill-posed.

There are numerous estimators of N_s that have been proposed in the literature, ranging from parametric to nonparametric, and including both frequentist and Bayesian. A common method, attributed to Good [183], is to take the observed number of species, say n_{obs}, and adjust it upward by a certain factor,

$$\hat{N}_s^{(cov)} = \frac{n_s^{obs}}{\hat{c}} \quad , \tag{5.11}$$

where $\hat{c} = 1 - x_1/n$, for x_1 equal to the number of species observed only once in the sample of size n. The estimator $\hat{N}_s^{(cov)}$ is a so-called *coverage estimator*, because the factor $\hat{c}$ is an estimate of the coverage c of the sample – that is, of the fraction of the population corresponding to those species observed at least once. This is a nonparametric estimator, but it has asymptotic behavior quite close to that of the maximum likelihood estimator, and it is generally much easier than the latter to compute. It has been noted, however, that in small samples these estimators can suffer from significant bias and large variance. Relevant details may be found in Darroch and Ratcliff [108] or Esty [138], for example.

We will see in Section 5.5 an example of a network species problem in the context of the Internet and a corresponding nonparametric estimator similar in spirit to (5.11).

5.3 Common Network Graph Sampling Designs

We here introduce some common designs for sampling a population network graph G. For convenience, our treatment is entirely in terms of undirected graphs, but most of the expressions can be generalized in a straightforward fashion to the case of directed graphs.

Graph sampling designs are somewhat distinct from typical sampling designs in non-network contexts (such as those seen above), in that there are effectively two inter-related sets of units being sampled, vertices i and edges $\{i, j\}$. Often these designs can be characterized as having two stages: a selection stage, followed by an observation stage. Selection is generally made among one class of units (e.g., vertices), which then leads to observation of units from the other class (e.g., edges) or even both classes. We will derive vertex and edge inclusion probabilities for the designs introduced below. These probabilities will be valuable both in lending insight into the nature of the designs and in providing us with a central element necessary to define Horvitz-Thompson estimators for certain key classes of network graph characteristics $\eta(G)$ in the next section.

5.3.1 Induced and Incident Subgraph Sampling

We first consider *induced subgraph sampling*, which consists of taking a random sample of vertices in a graph G and observing their induced subgraph. More precisely, in this design a simple random sample of n vertices is selected from V, without replacement, which yields the set $V^* = \{i_1, \ldots, i_n\}$. Edges are then observed for all vertex pairs $i, j \in V^*$ for which $\{i, j\} \in E$, yielding the set E^*. Note that this same design was encountered already in Example 5.1. A schematic illustration is shown in Figure 5.2. This type of sampling is representative of, say, the construction of contact networks in social network research, when a sample of individuals is first

selected and then the individuals are interviewed regarding some measure of contact among themselves (e.g., friendship, likes or dislikes, etc.).

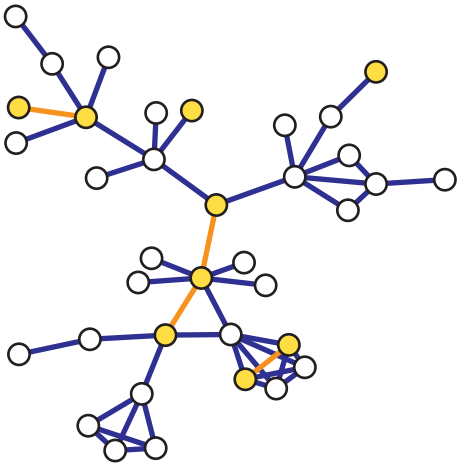

Fig. 5.2 Schematic illustration of induced subgraph sampling. Selected nodes are shown in yellow, while observed edges are shown in orange.

For induced subgraph sampling, the vertex and edge inclusion probabilities are uniformly equal to

$$\pi_i = \frac{n}{N_v} \quad \text{and} \quad \pi_{\{i,j\}} = \frac{n(n-1)}{N_v(N_v-1)} \ , \tag{5.12}$$

for all $i \in V$ and all $(i,j) \in V^{(2)}$, where $V^{(2)}$ is the set of all unordered pairs of vertices. These expressions follow from exactly the same arguments as underlie the calculation of the probabilities π_i and π_{ij} in Example 5.2. Note that they require knowledge of N_v. In the case of sampling a contact network, for example, this value would likely be available from database records if the overall population consisted of, say, all employees of a corporate entity.

Complementary to induced subgraph sampling is *incident subgraph sampling*. Instead of selecting n vertices in the initial stage, n edges are selected, again through simple random sampling without replacement, directly yielding the set E^*. All vertices incident to the selected edges are then observed, thus providing V^*. This design is illustrated in Figure 5.3. Such a design is, for example, implicit in the construction of sampled telephone call graphs, wherein telephone calls are sampled from a database, after which the phone numbers of the initiator and the receiver of the call are observed.

Regarding the inclusion probabilities for incident subgraph sampling, clearly the edge inclusion probabilities are just $\pi_{\{i,j\}} = n/N_e$. The form of the vertex inclusion probabilities, however, is more complicated, due to the fact that a vertex is included

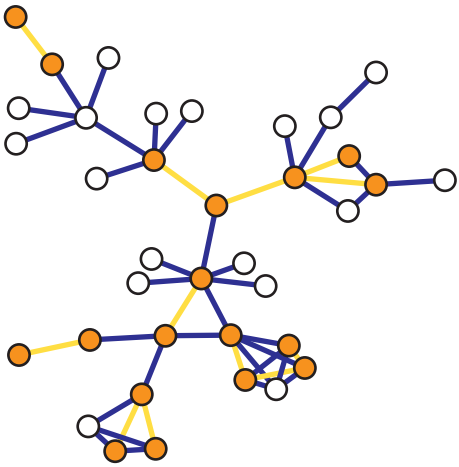

Fig. 5.3 Schematic illustration of incident subgraph sampling. Selected edges are shown in yellow, while observed nodes are shown in orange.

whenever any one or more of its incident edges is sampled. Specifically, we have

$$\begin{aligned}\pi_i &= \mathbb{P}(\text{ vertex } i \text{ is sampled }) \\ &= 1 - \mathbb{P}(\text{ no edge incident to } i \text{ is sampled }) \\ &= \begin{cases} 1 - \frac{\binom{N_e - d_i}{n}}{\binom{N_e}{n}}, & \text{if } n \le N_e - d_i \ , \\ 1, & \text{if } n > N_e - d_i \ , \end{cases}\end{aligned} \tag{5.13}$$

where d_i is the degree of vertex i.

Hence, in incident subgraph sampling, while the edges are included in the sample graph G^* with equal probability, the vertices are included with unequal probabilities that depend on the degree sequence. Looked at another way, incident subgraph sampling induces probability proportional to size sampling of vertices, reminiscent of Example 5.3, where 'size' here is driven by vertex degree. Clearly, to compute these inclusion probabilities, knowledge of both the number of edges N_e and the degrees of the nodes must be available. In the example of sampling a telephone call graph, for instance, this would require having access to marginal summaries of the total number of calls (say, in a given month) as well as the number of calls in which a given phone number had participated.

5.3.2 Star and Snowball Sampling

Another possible design is *star sampling*. In star sampling, an initial vertex sample V_0^* of size n is taken. Then all edges incident to vertices $i \in V_0^*$ are observed, yielding

E^*, and thus $G^* = (V^*, E^*)$, with $V^* = V_0^*$. Additionally, it may be that it is possible to observe not only these incident edges, but also the vertices $i \in V \setminus V_0^*$ to which these edges are also incident, thus enlarging the set V^* accordingly. These two variations have been referred to as *unlabeled* and *labeled* star sampling, respectively. In conducting a study of co-authorship among people within a given scientific literature, for example, randomly sampling records of n authors and recording the total number of co-authors of each author would correspond to unlabeled star sampling; if not only the number but the identities of the co-authors are recorded, this would correspond to labeled star sampling. A variant of unlabeled star sampling was also seen already in Example 5.1.

The vertex inclusion probabilities under unlabeled star sampling are just $\pi_i = n/N_v$, if the initial vertex set V_0^* is selected as a simple random sample without replacement. The edge inclusion probabilities, using an argument similar to that underlying (5.13), take the form

$$\begin{aligned} \pi_{\{i,j\}} &= \mathbb{P}(\text{ neither } i \text{ nor } j \text{ are sampled }) \\ &= 1 - \frac{\binom{N_v-2}{n}}{\binom{N_v}{n}}, \end{aligned} \tag{5.14}$$

which is roughly equal to $1-(1-n/N_v)^2$ for large n and N_v.

Note that the expression for the edge inclusion probabilities remains unchanged under the labeled version of star sampling. However, the vertex probabilities are indeed affected and can be shown to look like

$$\pi_i = \sum_{L \subseteq \mathcal{N}_i^+} (-1)^{|L|+1} \mathbb{P}(L) \ , \tag{5.15}$$

where $\mathcal{N}_i^+$ is the union of vertex i and those vertices that are immediate neighbors, $|L|$ is the cardinality of a given set L, and $\mathbb{P}(L)$ is the probability of selecting the set L in obtaining S. For example, under simple random sampling without replacement, we have

$$\mathbb{P}(L) = \frac{\binom{N_v-|L|}{n-|L|}}{\binom{N_v}{n}} \ . \tag{5.16}$$

The calculation of probabilities like those in (5.15) and (5.16) is simplified somewhat if we model the selection of vertices for the initial vertex subset V_0^* as arising through Bernoulli sampling. That is, the presence or absence of each vertex $i \in V$ in the sample V_0^* is held to be determined by an independent coin flip with probability p. In the case where N_v is large and $p \approx n/N_v$ is small, this Bernoulli sampling design is a quite reasonable approximation to simple random sampling without replacement.

An extension of the idea of star sampling is *snowball sampling* Specifically, whereas (labeled) star sampling extends an initial vertex sample V_0^* to only those vertices that are immediate neighbors of these, a snowball sample iterates this process in a natural manner. Define $\mathcal{N}(S)$ to be the set of all neighbors of vertices in a

set S. Then snowball sampling extends V_0^* to $V_1^* = \mathcal{N}(V_0^*) \cap \bar{V}_0^*$, and extends V_1^* to $V_2^* = \mathcal{N}(V_1^*) \cap \bar{V}_0^* \cap \bar{V}_1^*$, and so on. The set V_k^* is called the k-th *wave* of the sampling process. Sampling can be continued until a wave V_k^* is reached that is empty, or it can be stopped after some number K stages. Star sampling corresponds to the case $K = 1$. The final graph G^* obtained through snowball sampling consists of the vertices in $V^* = V_0^* \cup V_1^* \cup \cdots \cup V_K^*$ and, by construction, their incident edges.

An illustration of two-stage snowball sampling is shown in Figure 5.4. Some surveys of the World Wide Web graph can be thought of as arising through a variant of snowball sampling. Computer programs called 'spiders' are written to follow an initially compiled subset V_0^* of web pages to those pages corresponding to the HTML addresses listed in the V_0^*-pages. The newly discovered pages constitute S_1, and these in turn are then examined for new HTML addresses, which are then pursued. In other words, the 'spiders' mimic a human using a web browser by exhaustively following hyperlinks on discovered web pages.[2]

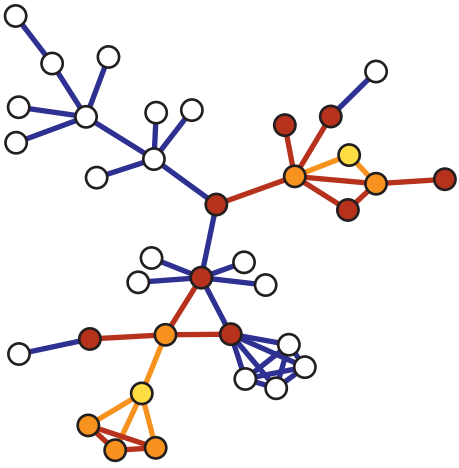

Fig. 5.4 Schematic illustration of two-stage snowball sampling. Nodes selected in the initial sampling are shown in yellow, while edges and nodes observed in the first and second waves of sampling are shown in orange and brown, respectively.

Unfortunately, although not surprisingly, inclusion probabilities for snowball sampling become increasingly intractable to calculate after the one-stage level corresponding to star sampling. The underlying formulas follow in principle from application of inclusion-exclusion arguments, but these lead to combinatorial issues that must be faced. A useful resource in this regard is Frank [148], although the treatment therein is still almost exclusively that of one-stage snowball sampling.

[2] Hyperlinks are directional, with one web page referencing another, and thus the web graph is more properly considered as a directed graph, a detail we have ignored here for the sake of exposition.

5.3.3 Link Tracing

There are various other designs for sampling network graphs. Many of them fall under the general label of *link-tracing* designs, in which after the selection of an initial sample, some subset of the edges ('links') from vertices in this sample are traced to additional vertices. Snowball sampling is a special case of link-tracing, in that all edges are followed until termination. But often it is not feasible, nor perhaps even desirable, to follow all edges. For example, in sampling social contact networks, it may be that individuals are unaware of or cannot recall all of their contacts, or that they do not wish to divulge some of them.

The inclusion probabilities under link-sampling of course vary with the design. As an illustration, consider the following idealized version of the `traceroute` sampling underlying Internet topology surveys like that described in Section 3.5.2. A sample $S = \{s_1, \ldots, s_{n_s}\}$ of n_s 'sources' are selected from the vertex set V of an Internet network graph G. Then a sample $T = \{t_1, \ldots, t_{n_t}\}$ of n_t 'targets' are selected from $V \setminus S$. Finally, a route is traced from each source node in S to each target node in T. That is, effectively a path is sampled between each pair $(s_i, t_j) \in S \times T$, and all vertices and edges in the paths are observed. The sampled graph $G^* = (V^*, E^*)$ is then constructed as the union of vertices and edges over all sampled paths. An illustration is shown in Figure 5.5.

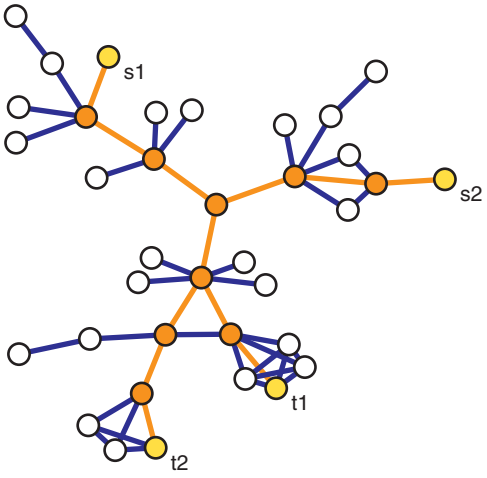

Fig. 5.5 Schematic illustration of the `traceroute` version of link-tracing. Selected source nodes $\{s_1, s_2\}$ and target nodes $\{t_1, t_2\}$ are shown in yellow, while nodes and edges observed on traces from sources to targets are shown in orange.

If the source and target sets are assumed to be obtained through simple random sampling without replacement, and if it is assumed that the corresponding paths sampled are shortest paths with respect to some set of edge weights, then it is argued in Dall'Asta et al. [107] that the vertex and edge inclusion probabilities behave like

$$\pi_i \approx 1 - (1 - \rho_s - \rho_t)\exp(-\rho_s \rho_t b_i) \tag{5.17}$$

and

$$\pi_{\{i,j\}} \approx 1 - \exp(-\rho_s \rho_t b_{i,j}) \ , \tag{5.18}$$

where b_i is the vertex betweenness centrality of i, $b_{i,j}$ is the edge betweenness centrality of edge $\{i,j\}$, and $\rho_s = n_s/N_v$ and $\rho_t = n_t/N_v$ are the source and target sampling fractions, respectively. From these expressions we see that `traceroute` sampling induces a type of probability proportional to size sampling of vertices and edges, where 'size' varies with the vertex and edge betweenness centralities.

Although the probabilities in (5.17) and (5.18) are not calculable from the `traceroute` observations themselves, they lend interesting insight into the nature of this sampling design, which we will return to in Section 5.5.

5.4 Estimation of Totals in Network Graphs

With appropriate choice of a population of units $\mathscr{U}$ and unit values y, various graph summary characteristics $\eta(G)$ encountered in Chapter 4 can be written in a form that involves a total $\tau = \sum_{i \in \mathscr{U}} y_i$. Examples include the following.

- Let $\mathscr{U} = V$ and $y_i = d_i$. The average degree of a graph G is obtained by scaling the total $\sum_{i \in V} d_i$ by N_v .
- Let $\mathscr{U} = E$ and $y_{\{i,j\}} = 1$. The number of edges N_e is given by the total $\sum_{\{i,j\} \in E} 1$.
- Let $\mathscr{U} = V^{(2)}$ and $y_{(i,j)} = I_{k \in \mathscr{P}_{(i,j)}}$ be the indicator of the event that the shortest path between i and j contains node k. In the case of unique shortest paths, the betweenness centrality $c_B(k)$ of a vertex $k \in V$ is given by the total $\sum_{(i,j) \in V^{(2)}} I_{k \in \mathscr{P}_{(i,j)}}$.
- Let $\mathscr{U} = V^{(3)}$ be the set of all triples of distinct vertices (i,j,k). The clustering coefficient $\mathsf{cl}_T(G)$ – that is, the transitivity of the graph G – can be written as three times the ratio of the total number of triples that form triangles to the total number of connected triples.

Often totals such as these can be estimated from a sampled graph G^* using generalizations of the Horvitz-Thompson estimator in (5.3), some instances of which we describe now.

5.4.1 Vertex Totals

As a starting point, consider the case of a *vertex total* (i.e., a total of the form $\tau = \sum_{i \in V} y_i$, for some choice of y). The first example above, involving the average degree, may be viewed as a vertex total. Another example is when y is a binary

variable indicating that a vertex has a given characteristic, in which case τ counts the number of vertices with that characteristic, and τ/N_v, the proportion. Quantities of this sort might include the fraction of web pages on the World Wide Web with links pointing to the Google homepage or the number of genes in a gene regulatory network responsible for the growth of an organism.

Given a sample of vertices $V^* \subseteq V$, we see from (5.3) that the Horvitz-Thompson estimator for vertex totals takes the form

$$\hat{\tau}_\pi = \sum_{i \in V^*} \frac{y_i}{\pi_i} \ , \tag{5.19}$$

where the π_i are the vertex inclusion probabilities corresponding to the underlying network sampling design. The variance of $\hat{\tau}_\pi$ in (5.19) and an unbiased estimate of that variance similarly follow directly from (5.5) and (5.6), with $\mathscr{U}$ and S replaced by V and V^*, respectively.

Note that in some cases the role of the network will be irrelevant for estimating a vertex total, such as when the characteristic y has nothing to do with the relational measurements defining the network and vertices are sampled through simple random sampling without replacement. In such cases, the estimator in (5.19) is just the conventional Horvitz-Thompson estimator. On the other hand, the network would be relevant if snowball sampling were used, since in that case the network structure itself plays a key role in the sampling of V^* and hence the calculation of the π_i.

5.4.2 Totals on Vertex Pairs

Now suppose that a quantity y_{ij} corresponding to vertex pairs $(i, j) \in V^{(2)}$ is of interest. Totals of the form

$$\tau = \sum_{(i,j) \in V^{(2)}} y_{ij} \tag{5.20}$$

are then relevant. The second and third examples at the start of this section, involving the number of edges N_e and the vertex betweenness c_B, are examples of this type of total. Examples of other quantities involving such totals include the total number of adjacent vertices with mutual labels (e.g., shared gender in a friendship network) and the average over the graph of some sort of (dis)similarity value between vertex pairs.

The Horvitz-Thompson estimator in this context takes the form

$$\hat{\tau}_\pi = \sum_{(i,j) \in V^{*(2)}} \frac{y_{ij}}{\pi_{ij}} \tag{5.21}$$

for estimating the total τ in (5.20). In the case that y_{ij} defining τ are nonzero only for vertex pairs (i, j) corresponding to edges $\{i, j\} \in E$, τ is an *edge total,* and

the inclusion probabilities in (5.21) are just the edge inclusion probabilities (i.e., $\pi_{ij} = \pi_{\{i,j\}}$).

Generalizing (5.5), the variance of the above estimator is given by

$$\mathbb{V}(\hat{\tau}_\pi) = \sum_{(i,j)\in V^{(2)}} \sum_{(k,l)\in V^{(2)}} y_{ij} y_{kl} \left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - 1 \right) \quad , \tag{5.22}$$

where π_{ijkl} is the probability that vertices i, j, k, and l are all included in the sample, and $\pi_{ijkl} = \pi_{ij}$ for convenience when $(i, j) = (k, l)$. The corresponding unbiased estimate of variance is given by

$$\widehat{\mathbb{V}}(\hat{\tau}_\pi) = \sum_{(i,j)\in V^{*(2)}} \sum_{(k,l)\in V^{*(2)}} y_{ij} y_{kl} \left(\frac{1}{\pi_{ij}\pi_{kl}} - \frac{1}{\pi_{ijkl}} \right) \quad . \tag{5.23}$$

Note that these quantities can become increasingly complicated to compute under some sampling designs, since it is necessary to be able to evaluate probabilities π_{ijkl} that four-tuples of vertices are sampled.

Example 5.4 (Estimating the Size of a Network Graph). Consider the problem of estimating N_e as an edge total, i.e.,

$$N_e = \sum_{\{i,j\}\in E} 1 = \sum_{(i,j)\in V^{(2)}} I_{\{i,j\}\in E} \quad . \tag{5.24}$$

Frank [149] derives explicit formulas for the Horvitz-Thompson estimator, its variance, and an unbiased estimate of this variance, in the context of induced subgraph sampling, where the vertices are sampled either with replacement, without replacement, or through Bernoulli sampling. For such designs – which Frank terms *symmetric sampling designs* – the quadruple vertex inclusion probabilities π_{ijkl} are functions only of the number of different vertices among i, j, k, and l, and not of the actual vertices themselves. That is, if there are $1 \le r \le 4$ different vertices among i, j, k, and l, then $\pi_{ijkl} = p_r$, where

$$p_r = \binom{n}{r} \Big/ \binom{N_v}{r} \tag{5.25}$$

for sampling n vertices without replacement,

$$p_r = \sum_{m=0}^{r} (-1)^m \binom{r}{m} (1 - m/N_v)^n \tag{5.26}$$

for sampling with replacement, and $p_r = p^r$ for Bernoulli sampling with probability p.

The estimator (5.21) of N_e in this situation is just

$$\hat{N}_e = \sum_{(i,j)\in V^{*(2)}} \pi_{ij}^{-1} = p_2^{-1} N_e^* \quad , \tag{5.27}$$

which simply scales up the empirically observed total N_e^* by a factor p_2^{-1}. The variance (5.22) takes the form

$$\mathbb{V}(\hat{N}_e) = \alpha_0 N_e^2 + \alpha_1 \mathscr{Q} + \alpha_2 N_e \ , \tag{5.28}$$

where $\mathscr{Q} = \sum_{i \in V} d_i^2$ is the sum of squares of the vertex degrees in G and

$$\begin{aligned} \alpha_0 &= (p_4 - p_2^2)/p_2^2 \\ \alpha_1 &= (p_3 - p_4)/p_2^2 \\ \alpha_2 &= (p_2 - 2p_3 + p_4)/p_2^2 \ . \end{aligned} \tag{5.29}$$

The variance estimator in (5.23) becomes

$$\widehat{\mathbb{V}}(\hat{N}_e) = \beta_0 N_e^{*^2} + \beta_1 \mathscr{Q}^* + \beta_2 N_e^* \ , \tag{5.30}$$

where $\mathscr{Q}^*$ is the analogue of $\mathscr{Q}$ on G^* and

$$\begin{aligned} \beta_0 &= \frac{1}{p_2^2} - \frac{1}{p_4} \\ \beta_1 &= \frac{1}{p_4} - \frac{1}{p_3} \\ \beta_2 &= \frac{2}{p_3} - \frac{1}{p_2} - \frac{1}{p_4} \ . \end{aligned} \tag{5.31}$$

Figure 5.6 shows the results of a numerical simulation, in which the true graph G is the network of protein interactions in yeast from Section 4.2.1.1, as in Example 5.1. Induced subgraph sampling was simulated in each of $10,000$ trials, using Bernoulli sampling of vertices with $p = 0.10, 0.20$, and 0.30. Shown in the figure are histograms of the estimators $\hat{N}_e$ in (5.27), for each choice of p, and of the estimated standard errors of these estimators, based on (5.30). The average of $\hat{N}_e$ over the simulations was 31116, 31197, and 31203, for $p = 0.10, 0.20$, and 0.30, respectively. Thus the unbiasedness of all three estimators in estimating $N_e = 31,201$ is well-supported by these results. The unbiasedness of the estimated variances (5.30) was similarly supported. Note that we do not expect unbiasedness of the estimated standard error, as the latter is a nonlinear function of the estimated variance. However, the average of these values is in fact reasonably close to the true underlying values. Note too, from the figure, that not only does the mean estimated standard error decrease with p, but so too does the variability of this quantity. □

Frank [151] offers results similar to those above for the estimation of certain totals in directed graphs.

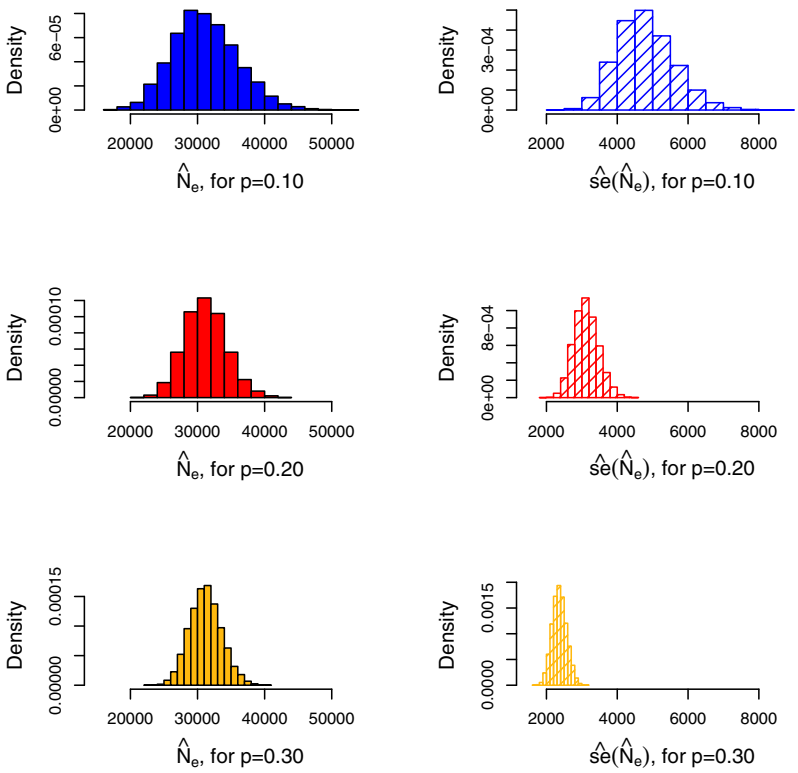

Fig. 5.6 Histograms of estimates $\hat{N}_e$ (left) of $N_e = 31,201$, as well as estimated standard errors (right), in the yeast protein interaction network, under induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.10$ (blue), 0.20 (red), and 0.30 (yellow). Results based on $10,000$ trials.

5.4.3 Totals of Higher Order

Having considered estimation of totals on vertices and vertex pairs, it should be clear that we can in principle continue in this manner and tackle totals on subsets of vertices of arbitrarily higher order. In practice, however, the relevant expressions become quite complicated. Triples of vertices appear to be both the order of next most natural interest beyond vertex pairs and the only other order studied in any detail.

For the case of totals $\tau = \sum_{(i,j,k)\in V^{(3)}} y_{ijk}$ indexed by triples of vertices, similar reasoning to that above leads to the Horvitz-Thompson estimator

$$\hat{\tau}_\pi = \sum_{(i,j,k)\in V^{*(3)}} \frac{y_{ijk}}{\pi_{ijk}} \quad . \tag{5.32}$$

Accompanying expressions for variance and estimated variance follow in a like manner, but we will not state them here. However, the following example provides some insight into these quantities.

Example 5.5 (Estimation of cl_T). Recall the definition of the clustering coefficient cl_T of a graph G, as given in (4.13). This quantity may be re-expressed in the form

$$\mathsf{cl}_T(G) = \frac{3\tau_\triangle(G)}{\tau_3^\dagger(G) + 3\tau_\triangle(G)} \quad , \tag{5.33}$$

where $\tau_3^\dagger(G) = \tau_3(G) - 3\tau_\triangle(G)$ is defined to be the number of vertex triples that are connected by exactly two edges. Thus $\mathsf{cl}_T(G)$ is a function of two different totals of the form $\sum_{(i,j,k)\in V^{(3)}} y_{ijk}$, where

$$y_{ijk} = A_{ij}A_{jk}A_{ki} \tag{5.34}$$

in the case of $\tau_\triangle(G)$ and

$$y_{ijk} = A_{ij}A_{jk}(1-A_{ki}) + A_{ij}(1-A_{jk})A_{ki} + (1-A_{ij})A_{jk}A_{ki} \tag{5.35}$$

in the case of $\tau_3^\dagger(G)$, and A is the adjacency matrix of G.

The network of protein interactions in yeast has $\tau_\triangle(G) = 44,858$ triangles, $\tau_3^\dagger(G) = 1,006,575$ triples connected by exactly two edges, and a clustering coefficient $\mathsf{cl}_T(G) = 0.1179$. We simulated 10,000 trials of induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.20$. Unbiased estimates of the two totals were obtained through the formulas $\hat{\tau}_\triangle(G) = p^{-3}\tau_\triangle(G^*)$ and $\hat{\tau}_3^\dagger(G) = p^{-3}\tau_3^\dagger(G^*)$. Histograms of the resulting values are shown in Figure 5.7. The average value for these estimators over the 10,000 trials was 44,681 and 1,004,963, respectively, while their standard errors were 12,425 and 245,051. The results therefore support the unbiasedness of the estimators, but suggest that both are rather variable.

A plug-in estimator of $\mathsf{cl}_T(G)$ follows by the appropriate substitution of $\hat{\tau}_\triangle(G)$ and $\hat{\tau}_3^\dagger(G)$ in (5.33). Note that the resulting value $\widehat{\mathsf{cl}}_T$ is just $\mathsf{cl}_T(G^*)$. We see that under the sampling design used here, this estimator appears to perform reasonably well, with an average value of 0.1191 and a standard error of 0.0251 over the 10,000 trials. □

Results for estimation of analogous totals on vertex triples in directed graphs may be found in Frank [151], for the case of induced subgraph sampling, with simple random sampling of vertices without replacement.

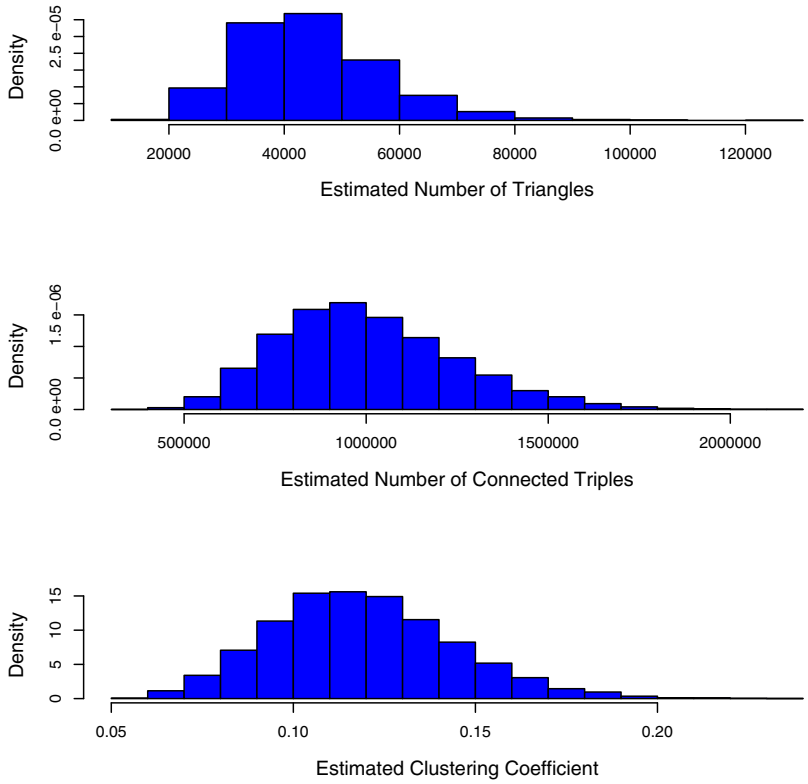

Fig. 5.7 Histograms of estimates $\hat{\tau}_{\triangle}(G)$ (top), $\hat{\tau}_3^{\dagger}(G)$ (middle), and $\widehat{\mathrm{cl}}_T(G)$ (bottom) in the yeast protein interaction network, under induced subgraph sampling, with Bernoulli sampling of vertices, using $p = 0.20$. True values being estimated were $\tau_{\triangle}(G) = 44,858$, $\tau_3^{\dagger}(G) = 1,006,575$, and $\mathrm{cl}_T(G) = 0.1179$. Results based on 10,000 trials.

5.4.4 Effects of Design, Measurement, and Total

On a final note, although the above-described framework for estimation of various totals in network graphs is fairly straightforward in its essence, it is important to realize that the interaction of the sampling design, the measurements taken, and the total to be estimated has a direct effect on whether and how this framework may be used effectively. In particular, three elements must be present in the problem at hand. First, the network graph summary statistic $\eta(G)$ must, of course, be expressed in terms of a total(s). Second, the values y underlying the definition of this total must be either observed or obtainable from the available measurements. Third, the relevant inclusion probabilities π must be computable for the underlying network

sampling design. Unfortunately, it is not always the case that all three elements are present at the same time. The following example helps illustrate this point.

Example 5.6 (Estimating Average Degree (Continued)). We consider again the problem of estimating the average degree of a network graph G (i.e., $= \tau/N_v$, where $\tau = \sum_{i \in V} d_i$). Also, as in Example 5.1, we again consider two possible network sampling designs – unlabeled star sampling and induced subgraph sampling. We write the resulting sampled graphs as $G^*_{St} = (V^*_{St}, E^*_{St})$ and $G^*_{IS} = (V^*_{IS}, E^*_{IS})$, respectively. Recall that under both designs vertices are selected through simple random sampling without replacement from V, but that all edges adjacent to vertices $i \in V^*_{St}$ are observed in forming E^*_{St}, whereas only edges between pairs of vertices $i, j \in V^*_{IS}$ are observed in forming E^*_{IS}.

The average degree is clearly a rescaling of a vertex total. Therefore, it seems natural to use the estimator in (5.19) here. Implicitly, this approach requires that $y_i = d_i$ be observed. Under the star sampling design this is indeed the case, and we arrive at the estimator

$$\hat{}_{St} = \frac{\hat{\tau}_{St}}{N_v} \quad , \quad \text{where} \quad \hat{\tau}_{St} = \sum_{i \in V^*_{St}} \frac{d_i}{n/N_v} \quad . \tag{5.36}$$

On the other hand, under induced subgraph sampling, we do not observe d_i. Rather, as remarked in Example 5.1, we effectively observe a number $d^*_i \leq d_i$ for each vertex $i \in V^*_{IS}$. As a result, τ is not amenable to Horvitz-Thompson estimation methods as a vertex total under this design.

Note, however, the relation $= 2N_e/N_v$ shows that estimation of may alternatively be approached as a variation on the problem of estimating the graph size N_e. And under induced subgraph sampling, as we saw above in Example 5.4, we can estimate N_e in an unbiased fashion by

$$\hat{N}_{e,IS} = \sum_{\{i,j\} \in E^*_{IS}} \frac{1}{n(n-1)/[N_v(N_v-1)]} = N^*_{e,IS} \cdot \frac{N_v(N_v-1)}{n(n-1)} \quad , \tag{5.37}$$

which yields the unbiased estimator

$$\hat{}_{IS} = \frac{2\hat{N}_{e,IS}}{N_v} \tag{5.38}$$

for .

The estimators (5.36) and (5.38) may be usefully compared by re-writing them as

$$\hat{}_{St} = \frac{2N^*_{e,St}}{n} \quad \text{and} \quad \hat{}_{IS} = \frac{2N^*_{e,IS}}{n} \cdot \frac{N_v - 1}{n-1} \quad . \tag{5.39}$$

We see that under star sampling our estimator simply results from the use of the formula $2N_e/N_v$ with the graph G^*_{St}, since $N^*_{St} = n$. In contrast, under induced subgraph sampling, the analogous quantity is inflated by the factor $(N_v - 1)/(n-1)$, which

seeks to correct for the extent to which the degrees d_i are 'thinned' in observing the d_i^*. □

5.5 Estimation of Network Group Size

Throughout most of the previous section, we implicitly assumed that N_v was known. In some settings, this may not be the case. Some groups, like many human and animal populations, are too mobile or elusive to count accurately. Others, like the collection of all web pages or all Internet routing devices, are too massive and dispersed to survey in their entirety. In these settings, estimation of N_v itself may be a prime objective.

If it is feasible to use a network sampling design that selects vertices through simple random sampling without replacement or Bernoulli sampling, then we may entertain the idea of doing so twice, after 'marking' the first sample, and using capture-recapture estimators like that in (5.9) directly 'off the shelf.' When it is not feasible to select vertices in this manner, or when the basic capture-recapture estimator performs poorly, it may be necessary or at least desirable to develop analogous estimators of N_v explicitly tailored to the sampling and network at hand. We illustrate with two examples.

Example 5.7 (Estimating the Size of a 'Hidden Population'). The term 'hidden population' generally refers to one in which the individuals do not wish to expose themselves to view. For example, humans of socially sensitive status, such as the homeless, or who are involved in socially sensitive activities, such as illegal drug usage or prostitution, will typically not be inclined to disclose themselves. This tendency, coupled with the fact that such subpopulations often are also quite small, makes estimating their size a particularly challenging problem. Simply sampling the overall population and using the observed proportion of hidden individuals to create an estimate of their number is likely to require an unacceptably large sample size for the estimate to have an acceptably small level of error.

As an alternative, Frank and Snijders [154] describe how snowball sampling may be used for this problem. In essence, the successive waves of the snowball sample provide information not unlike the re-observed 'marked' individuals in standard capture-recapture methods. Various estimators can be defined to exploit this information. We describe one such estimator to illustrate the underlying principle.

Let V be the set of all members of the hidden population, and let $G = (V, E)$ be a directed graph associated with that population, in which an arc from vertex i to vertex j indicates that, if asked, individual i would mention individual j as a member of the hidden population.[3] Let G^* be a subgraph of G, where the vertices $V^* = V_0^* \cup V_1^*$ are obtained through a one-wave snowball sample, with the initial sample V_0^* selected through Bernoulli sampling from V with probability p_0.

[3] The practical concerns of trust, veracity, etc. associated with such questioning are themselves non-trivial issues of survey methodology, but we will ignore them here.

Our estimator of N_v will be derived using the method-of-moments. Let N, M_1, and M_2 be random variables defined such that $N = |V_0^*|$ is the size of the initial sample, M_1, the number of arcs among individuals in V_0^*, and M_2, the number of arcs pointing from individuals in V_0^* to individuals in V_1^*. The expectations of these three quantities, with respect to the randomness in the selection of the initial sample V_0^*, are given by

$$\mathbb{E}(N) = \mathbb{E}\left(\sum_i Z_i\right) = N_v p_0 \tag{5.40}$$

$$\mathbb{E}(M_1) = \mathbb{E}\left(\sum_{i \neq j} Z_i Z_j A_{ij}\right) = (N_e - N_v)\, p_0^2 \tag{5.41}$$

$$\mathbb{E}(M_2) = \mathbb{E}\left(\sum_{i \neq j} Z_i (1 - Z_j) A_{ij}\right) = (N_e - N_v)\, p_0 (1 - p_0) \quad , \tag{5.42}$$

where Z_i is a binary variable indicating whether vertex i is in the initial sample, and A_{ij} is the (i, j)-th entry of the adjacency matrix for G. Setting the left-hand sides of these equations equal to their observed counterparts, say n, m_1, and m_2, the first resulting equation yields the expression $N_v = n/p_0$, while the second two together yield the estimate $\hat{p}_0 = m_1/(m_1 + m_2)$ for the unknown value p_0. Upon substitution we obtain

$$\hat{N}_v = n\left(\frac{m_1 + m_2}{m_1}\right) \quad . \tag{5.43}$$

In other words, the number of individuals observed initially is inflated by an estimate of the sampling rate, where that estimate reflects the relative number of arcs from individuals in the initial sample that point inwards among themselves. It is interesting to compare the above estimator to that in (5.9).

Frank and Snijders [154] offer a number of other estimators of N_v in this spirit, using both design-based and model-based approaches. We note that the estimator in (5.43), while derived from a design-based perspective here, can also be derived from a model-based perspective. See Example 6.3 in Chapter 6.

Also proposed by these authors is an approach to estimating the variance of these estimators, based on the jackknife principle. Specifically, let $\hat{N}_{v,(i)}$ be the estimate of N_v obtained when vertex i is removed from the original sample V_0^*, as well as any vertex $j \in V_1^*$ having only i pointing to it among those vertices in V_0^*. Then the suggested estimate of variance is of the form

$$\widehat{\mathbb{V}}_J(\hat{N}_v) = \frac{n-2}{2n} \sum_{i \in V_0^*} \left(\hat{N}_{v,(i)} - \hat{N}_{v,(\cdot)}\right)^2 \quad , \tag{5.44}$$

where $\hat{N}_{v,(\cdot)}$ is the average of the estimates $\hat{N}_{v,(i)}$ over $i \in V_0^*$. It should be noted that the factor $(n-2)/(2n)$ in (5.44) differs from the factor $(n-1)/n$ typically found in jackknife estimates of variance for iid observations. The change can be justified as accounting for the effects of the fact that (i) there are essentially n^2 'observations' in

the $n \times n$ adjacency matrix, instead of just n, and (ii) these are not iid observations. See Frank and Snijders [154] for details. □

Example 5.8 (How Large Is the Internet?). The Internet is massive – and still growing. It is therefore of interest, from both practical and intellectual perspectives, to ask just how large the Internet actually is. Of course, there are various ways in which we might wish to define the 'size' of the Internet and various measurement techniques that might be used to obtain relevant data. Here we will simplify the problem to one of inferring the number of vertices N_v in an undirected graph G, based on a set of measurements taken according to the `traceroute` design described in Section 5.3. Our treatment follows that of Viger et al. [388].

Interestingly, the estimation of N_v from `traceroute` measurements can be viewed as a species problem. Recall that under our idealization of `traceroute` sampling the observed vertex set V^* in the sampled graph G^* is the set of all vertices discovered on paths from n_s sources $S = \{s_1, \ldots, s_{n_s}\}$ to n_t targets $T = \{t_1, \ldots, t_{n_t}\}$. A given vertex may be discovered on more than one path. Hence, the vertex may be thought of as a species, and each time that vertex is observed on a path, this may be viewed as having seen a member of that species.

Viger et al. [388] propose a solution to this particular species problem using a method based on principles of sample re-use. The basic idea is to ask whether a given target vertex $t_j \in T$, if dropped from the study, would have been discovered on paths to the other remaining target vertices. One can easily determine what fraction of targets would not have been discovered in this manner, and that fraction can be used to inflate the observed number of vertices N_v^* upward to create an estimate $\hat{N}_v$ of N_v, as we describe now.

Let $V^*_{(-j)}$ denote the number of vertices discovered on sampled paths to targets other than t_j, and define $\delta_j = I\left\{t_j \notin V^*_{(-j)}\right\}$ to be the indicator of the event that target t_j is *not* 'discovered' on sampled paths to any other target. Write the total number of such targets as $X = \sum_j \delta_j$.

Given a set of pre-selected source nodes (chosen either randomly or not), if the target nodes in T are chosen by simple random sampling without replacement from $V \setminus S$, the probability that target t_j is not discovered on the paths to other targets is given by

$$\mathbb{P}\left(\delta_j = 1 \,|\, V^*_{(-j)}\right) = \frac{N_v - N^*_{(-j)}}{N_v - n_s - n_t + 1} \quad , \tag{5.45}$$

where $N^*_{(-j)} = \left|V^*_{(-j)}\right|$. Note that, by symmetry under simple random sampling, the expectation $\mathbb{E}\left(N^*_{(-j)}\right)$ is the same for all j. We denote this quantity by $\mathbb{E}\left(N^*_{(-)}\right)$ and, as a result, we obtain

$$\mathbb{E}(X) = \sum_{j=1}^{n_t} \mathbb{P}\left(\delta_j = 1 \,|\, V^*_{(-j)}\right) = \frac{n_t\left[N_v - \mathbb{E}\left(N^*_{(-)}\right)\right]}{N_v - n_s - n_t + 1} \quad . \tag{5.46}$$

Rewriting this equation to isolate N_v, we have

$$N_v = \frac{n_t \mathbb{E}\left(N^*_{(-)}\right) - (n_s + n_t - 1)\mathbb{E}(X)}{n_t - \mathbb{E}(X)} \quad . \tag{5.47}$$

Estimation of N_v is thus reduced to estimation of $\mathbb{E}(X)$ and $\mathbb{E}(N^*_{(-)})$. Unbiased estimators of these two quantities are the value of X itself and the average of the $N^*_{(-j)}$, respectively. Under a slight variation of this idea and ignoring trivial terms and factors, Viger et al. [388] arrive at an estimator of the form

$$\hat{N}_v = (n_S + n_t) + \frac{N^*_v - (n_S + n_t)}{1 - w^*} \quad , \tag{5.48}$$

where $w^* = X/(n_t + 1)$. In other words, N_v is estimated by counting the source and target vertices, and then adjusting upwards the number of additional vertices discovered through `traceroute` sampling, by a factor reflecting the average tendency of targets to be discovered by other paths. This estimator has the advantage over the simple plug-in estimator suggested just above in that even if $X = n_t$ it remains well defined.[4]

A simple estimate of the variance of $\hat{N}_v$ in (5.48) is provided by the expression

$$\widehat{\mathbb{V}}(\hat{N}_v) \approx \frac{(N^*_v - n_S - n_t)^2 w^*}{(1 - w^*)^3 n_t} \quad , \tag{5.49}$$

which follows using a delta-method argument. See Viger et al. [388] for details.

Figure 5.8 illustrates the effectiveness of the estimator $\hat{N}_v$, as applied to a part of the Internet inferred by the Internet mapping project Skitter.[5] This network had $N_v = 624,324$ vertices and $N_e = 1,191,525$ edges. Sampling was simulated using $n_s = 10$ sources and a target sampling density $\rho_t = n_t/N_v$ that ranged from 0.001 to 1.0. The results show that the estimator can be quite accurate for even low target sampling rates and is certainly a vast improvement over the observed number of nodes N^*_v. □

The above are just two examples of group-size estimation problems arising in network contexts. It would appear that species problems are encountered often in this setting. For example, in Example 5.6, if we were forced to estimate the average degree, using techniques for vertex totals, under induced subgraph sampling, the dilemma faced there is essentially a species problem, in not knowing how many edges there are for each observed vertex $i \in V^*$. Similarly, in Example 5.8, estimation of the number of edges N_e is also a species problem, in that a given edge may be observed on multiple paths. Similarly, the estimation of the degree of a given node

[4] The derivation of the estimator requires an assumption equivalent to saying that the number of vertices uniquely discovered on paths to any one or any pair of target vertices is relatively small compared to the overall number of vertices N^*_v in the sampled graph V^*. This assumption appears to be well motivated by empirical findings in the literature.

[5] See *http://www.skitter.org*.

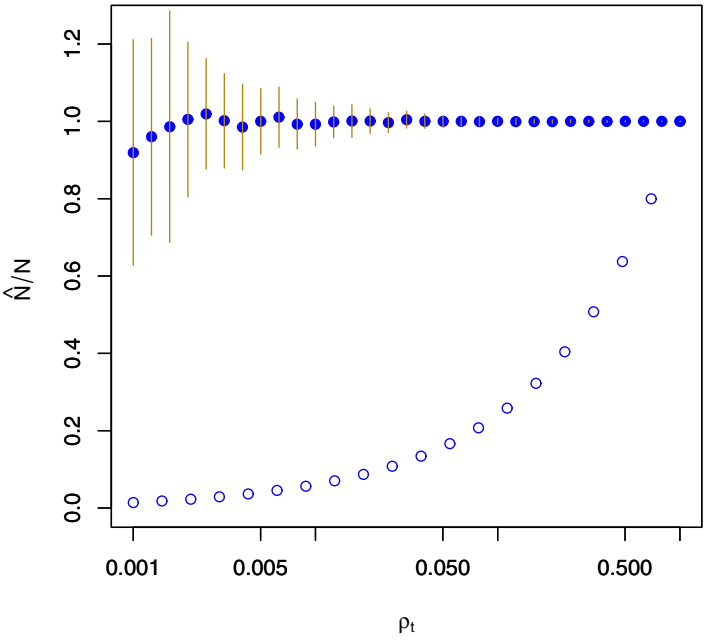

Fig. 5.8 Comparison of $\hat{N} = \hat{N}_v$ (filled circles) and $\hat{N} = N_v^*$ (open circles), as estimators of N_v, for various values of target sampling density ρ_t. Vertical bars indicate intervals of one standard deviation above and below each point, capturing the variation over multiple choices of source and target nodes. (Bars for the values N_v^* are too small to be visible.)

under `traceroute` sampling is a species problem too, as has been observed by Zhang [412].

5.6 Other Network Graph Estimation Problems

There are many other network graph estimation problems, some of which have been studied to an extent, and many of which have not. If a network graph G is sampled, rather than enumerated, and interest is in properties of G itself, then the computation of pretty much any characteristic $\eta(G)$ becomes a *de facto* estimation problem. However, not all are directly amenable to the types of methods we have described in this chapter, as the following example shows.

Example 5.9 (Estimation of Degree Frequency). A number of papers have demonstrated the effects that different network sampling designs can have in rendering

the observed degree distribution unrepresentative of the true underlying degree distribution, with particular interest in the issue of broad versus concentrated degree distributions. For instance, in Lakhina, Byers, Crovella, and Xie [243], results from numerical experiments are presented showing that extreme forms of `traceroute` sampling can actually induce a broad degree distribution in the sampled graph G^* when none exists in the true graph G. Analytical work (e.g., Clauset and Moore [91], Achlioptas, Clauset, Kempe, and Moore [3]) has confirmed and refined these findings, showing that even when G has an actual power-law degree distribution, the exponent α can be significantly under-estimated by exponents obtained from the degree distribution of G^*. Similar results have been demonstrated by Han et al. [192] in the case of sampling protein-protein interaction networks. See also Stumpf and Wiuf [373] and Stumpf, Wiuf, and May [374].

The degree distribution is susceptible to sampling effects in the same way that we have already seen the average degree to be; however, the design of accurate estimators for the average is substantially easier than for the full distribution. Let f_d and f_d^* be the true and observed frequencies of degree d nodes in G and G^*, respectively. Frank [153] shows using a combinatorial argument that, under induced subgraph sampling,

$$\mathbb{E}(f_d^*) = \sum_{d'=0}^{N_v-1} P(d,d') f_{d'} \quad , \tag{5.50}$$

where

$$P(d,d') = \frac{\binom{d'}{d}\binom{N_v-1-d'}{n-1-d}}{\binom{N_v-1}{n-1}} \quad . \tag{5.51}$$

In principle, pursuing a method-of-moments estimator, we can substitute f_d^* for $\mathbb{E}(f_d^*)$ on the left-hand side of (5.50), to obtain a system of equations for the vector $(f_0, \ldots, f_{N_v})$. But the resulting system will generally be under-determined, unless it is known *a priori* that the degree of G is bounded by a value $d_{max} \leq n$ and the equations are restricted accordingly. Furthermore, even in this setting, the solution to this system is not guaranteed to be non-negative. Finally, similar expressions for the variance of any such estimator would seem to be difficult to derive. □

There are other estimation problems in network graph sampling that have received attention, such as work by Frank [150] on the estimation of the number of connected components in a graph. Overall, however, while the literature on sampling and estimation in network graphs is nontrivial in its size and breadth, much of this material pre-dates the explosion of interest in networks of the past 5 to 10 years, and as a result there are many open problems in this area. Examples include the analysis of non-traditional sampling designs (e.g., particularly adaptive designs), the estimation of quantities not easily expressed as totals (e.g., degree distribution exponents), and the incorporation of effects of sampling error and missingness.

On a final note, we point out that sampling and estimation are also being used in a proactive manner in the context of large network graphs, as a way of producing computationally efficient 'approximations' to quantities that, if computed for the full network graph, would be prohibitively expensive. Examples include the estimation

of centrality measures (e.g., Eppstein and Wang [133], Brandes and Pich [57]), and the detection of so-called 'network motifs' in work of Kashtan, Itzkovitz, Milo, and Alon [218]. We will encounter the latter topic in Section 6.2.4.3.

5.7 Additional Related Topics and Reading

The elements of the material presented here go back to foundations laid in classical sampling theory, for which there are many good references, such as Thompson [379]. In the specific context of network graphs, a substantial fraction of the work done in this area is due to Ove Frank, appearing in a series of papers throughout the 1970's and 1980's. See Frank [152] for a recent overview and extensive bibliography. For background on the species problem, see the survey of Bunge and Fitzpatrick [69].

Exercises

5.1. Recall that for estimating a total τ, the Horvitz-Thompson estimator $\hat{\tau}_\pi$ in (5.3) is unbiased, in the sense that its expected value over all possible samples S is equal to τ. Show that the corresponding variance estimate $\widehat{\mathbb{V}}(\hat{\tau}_\pi)$ in (5.5) is an unbiased estimate of the true variance $\mathbb{V}(\hat{\tau}_\pi)$ in (5.5). That is, show that

$$\mathbb{E}\left(\widehat{\mathbb{V}}(\hat{\tau}_\pi)\right) = \mathbb{V}(\hat{\tau}_\pi) \ .$$

(Hint: You will want to introduce binary random variables $Z_1, \ldots, Z_{N_u}$, with $Z_i = 1$ indicating that unit i is in the sample S, as we did in Section 5.2 in showing the unbiasedness of $\hat{\tau}_\pi$ as an estimator of τ.)

5.2. Suppose that for every pair of vertices $i, j \in V$ in a graph G there is a unique shortest path, say $\mathscr{P}(i, j)$, between them. Then the betweenness centrality of a vertex $k \in V$ can be expressed as a total in the form

$$c_B(k) = \sum_{i \neq j \neq k \in V} I_{k \in \mathscr{P}(i,j)} \ .$$

Consider the problem of estimating $c_B(k)$, for fixed k.

a. If you obtain a sample of n vertices V^* by simple random sampling without replacement from $V \setminus \{k\}$, show that for each pair $i \neq j \in V \setminus \{k\}$ the probability of inclusion takes the form

$$\pi_{ij} = \frac{n(n-1)}{(N_v - 1)(N_v - 2)} \ .$$

b. For a network graph G of your choice, conduct a simulation study to evaluate the performance of the estimator

$$\hat{c}_B(k) = \sum_{(i,j)\in V^{*(2)}} \frac{I_{k\in\mathscr{P}(i,j)}}{\pi_{ij}} \quad ,$$

where the π_{ij} are as in part (a) and you assume the indicators $I_{k\in\mathscr{P}(i,j)}$ are observable for any pair $(i,j) \in V^{*(2)}$. Explore the characteristics of the estimators' performance as a function of the characteristics of the vertices k (e.g., degree, centrality c_B, etc.).

(Note: You may need to equip the edges e of your graph with weights w_e to enforce unique shortest paths. Weights that differ from one by small, random additive perturbations should be sufficient.)

5.3. Verify the moment equations in (5.40), (5.41), and (5.42). Show that the method-of-moments estimator of the number of vertices N_v yielded by these equations is given by the expression in (5.43).

5.4. Recall the discussion and references in the first paragraph of Example 5.9, regarding the extent to which the observed degree distributions $\{f_d^*\}$ of sampled graphs G^* have been found to deviate from the degree distributions $\{f_d\}$ of a true underlying network graph G.

a. Using a network graph G and a sampling design of your choice, explore this phenomenon using numerical simulation. To what extent do you find that the degree distributions of G^* and G differ? How sensitive are your results to tuneable aspects of your sampling design (e.g., the number n of vertices sampled)?

b. Under induced subgraph sampling, using your network graph G from part (a), explore the extent to which the proposed estimator deriving from (5.50) is feasible in your case. If it is not feasible – for example, if $d_{max} \gg n$ – can it be used to obtain accurate estimates of the distribution f_d up through some appropriately specified $d'_{max} \leq n$?