# Research Topic 2: Sampling and Estimation Methods for Partially Observed Networks.

*Author:* Zak Varty[†]

*Supervisor:* Dr. Christopher Nemeth[‡]

April 2017

[†] STOR-i Centre for Doctoral Training, Lancaster University.
[‡] Department of Mathematics and Statistics, Lancaster University.

## 1 Introduction and Motivation

Networks are a prolific part of modern life, appearing in some unexpected areas as well as those which are more commonly considered. The networks of interest to us are often very large, or else not easily accessible, as we will see in the following motivating examples.

Very large networks are abundant in modern life, and are frequently the subject of research. Social media, of course, with its massive popularity makes for a relevant and engaging set of human networks to study. Another obvious set of networks, this time physical, comes from the transportation industry, with road and rail networks providing large and interesting examples. The more abstract examples networks are perhaps less widely known, but nonetheless are fascinating topics of current research. Two examples of these abstract networks can be drawn from biology and linguistics; how and if the thousands of proteins in the body interact with one another, and the narrative connections between the subjects of a body of text.

Additionally, there are also many common examples of networks we would like to study but are not available to us. These include terrorist networks, which we wish to study in order to dismantle or disrupt most efficiently, and also networks of drug suppliers and users where we aim to best target crime interventions and health care services. Again there are many less obvious examples of these types of network, including populations who are not easily identifiable due to being stigmatised, marginalised or persecuted.

In each of these cases it is often not possible to observe the full network. In the former set of cases this is often because the full network is too large to handle or to collect data on, while in the latter it is because access to the full network is not possible. This leads to *partially observed networks*, and the need to estimate properties of the full network using only the partial information available. We shall see how more general statistical methods for sampling and making inference about the population as a whole may be adapted for use on partially observed networks, and some of the challenges this presents.

We begin with an brief background in the properties of graphs and three models frequently used for networks in Section 2. Sections 3 and 4 consider sampling schemes and estimation of population network properties for very large and hidden networks respectively. Section 5 then provides a review of these methods and highlights some current areas of research.

## 2 Background and Graph Terminology

Networks are usually represented mathematically as graphs. For this reason we spend this section looking briefly at some definitions of graph properties and common models for random graphs which will be useful in the study of networks. For a more complete background, see Salter-Townshend et al. [11] and Kolaczyk [10].

## Graphs and their properties

A *graph* $G = (V, E)$ is composed of a set of $N_v$ nodes, $V = \{v_1, \dots, v_{N_v}\}$, and a set of $N_e$ edges, $E = \{e_1, \dots, e_{N_e}\}$, which are unordered pairs of nodes $e_n = \{v_j, v_k\}$. Therefore $E \subseteq V^{(2)}$ where $V^{(2)}$ is the set of all unordered pairs of nodes. The edges correspond to links which are present between the nodes in the network, and although not considered in this report may be directed - an edge would then be specified by an ordered pair of nodes. The *size* of a graph is given by the number of edges it consists of, $|E|$, and the *order* of a graph is the number of nodes it is made up of,$|V|$. The nodes, edges or both component types of a graph may be *decorated* by having covariate information attached to them. An edge decoration in a rail network, for example, might be the distance between the stations it connects, while a node decoration might be the number of annual passengers through each station.

Rather than describing a graph by its node and edge sets, an adjacency matrix $Y$ could can also be used to fully describe a graph $G$. Two nodes are said to be *adjacent* if there is an edge which links them, and this edge is said to be *incident* to both nodes. The adjacency matrix $Y$ has as its entries $y_{ij}$ 1 if nodes $v_i$ and $v_j$ are adjacent and 0 otherwise. The *degree* of a node is the number of edges to which it is incident, and is therefore given by $\mathrm{d}_{v_i} = \sum_j y_{ij}$. A graph is said to be a *tree* if it is connected, so that there is a sequence of non-repeating edges between any two nodes, and that this sequence of edges between any two nodes is unique. A forest is then any graph which may is composed of the union of two or more trees.

## Common network models

The Erdos-Renyi model is the simplest and most studied probabilistic model for a network. It assumes that the edges, the $y_{ij}$, are independent and identically distributed Bernoulli($\theta$) random variables. Exponential random graph models extend this idea, adapting logistic regression to networks. In these models, the probability of observing a particular realisation of the adjacency matrix $y$ for a network is proportional to the exponent of some linear combination of a vector of network features, $S(y)$, where the features sufficiently describe the network.

$$P(Y = y) \propto \exp(\beta^T S(y))$$

The third network model mentioned is that of scale Free networks, which are networks in which the degree of the nodes may be described by a power law distribution,

$$P(\deg(v_i) = j) \propto j^{-\gamma} \text{ for some } \gamma.$$

Scale free networks are formed when nodes arrive sequentially and link to each of the other nodes with probability proportional to the number of links involving that node already [3].

# 3 Sampled Network Methods

While complete knowledge of graph structure is often assumed, this is not always a valid or good assumption. For small networks, such as social ties in a classroom or small office, full knowledge may be possible. This is not the case as the population size grows. In a large company it would not be reasonable to assume complete recall of interactions and it may in addition be prohibitively expensive to survey all members of the company. In a social media network, connections can be fully established easily and accurately, but it may not be possible to handle the entire network at once using the available computational power.

In this section we will consider circumstances where the full *population network*, for some reason, can not be used or gathered entirely and instead a *sampled network* is used. The sampled network $G^*$ is a subset of the nodes and edges in the population network $G$ , observed by following some sampling scheme. That is, $G^* = \{V^*, E^*\}$ is a subnetwork of $G = \{V, E\}$ where $V^* \subseteq V$ and $E^* \subseteq E$ are determined by the sampling scheme $\mathscr{S}$. The sampled network is then used to infer one or more characteristics of the full network $\eta(G)$. The way in which the sample is taken then effects how we should go about inferring properties of the population network as much as the properties of the population network do. As we shall see, many of the methods and issues in doing this are common to survey sampling design, though with the added complexity that comes with sampling components of a network rather than representatives of an unconnected population.

## 3.1 Sampling schemes for large networks

In this section we look at two possible sampling designs which could be used to obtain a sampled network from a population network of known order, $|V| = N_v$, or size $|E| = N_e$. We then consider how, under one of these sampling designs we might go about inferring a population network property $\eta(G)$ using the sampled network that is induced, $G^*$.

Sampling a network is distinguishable from sampling a population in that there are two distinct but related component types in a network, namely nodes and edges. This motivates graph sampling designs which are characterised by a section step and an observation step. A selection is made among one set of components and then components from the other set (or both sets) which are incident to the selection are also included.

### Induced subgraph sampling

Induced subgraph sampling first selects from among the set of nodes and then observes from the set of edges. Initially, a simple random sample without replacement of size $n$ is taken from the node set $V$, yielding the node set for the sampled graph $V^*$. The sampled edges are then observed as being those which pass between sampled nodes, $E^* = \{\{i, j\} \in E : i, j \in V^*\}$. In a social network example, this would be equivalent to forming your sample node set by taking a simple random sample without replacement from all people in the network and then observing the edge set to be connections between the selected individuals. Figure 1 illustrates induced subgraph sampling on a network of order 20.

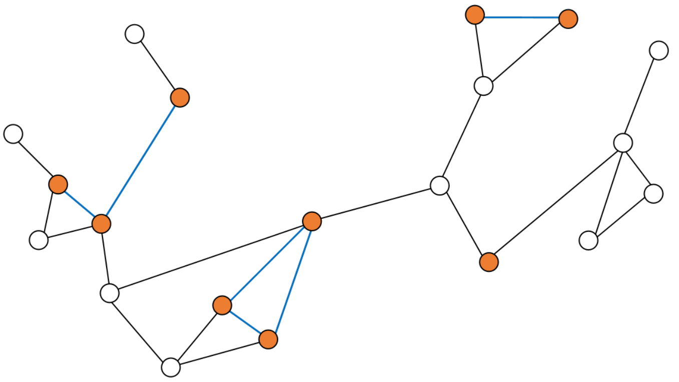

Figure 1: Induced subgraph sampling on a network of order 20. First the 9 nodes shown in orange are selected and then blue edges are observed to form the sampled network.

The probability of a component being included in the sampled network will be of interest when we consider Horvitz-Thompson estimators in Section 3.3, and so we note these here.

We begin by looking at the node inclusion probabilities. There are $\binom{N_v}{n}$ node samples $V^*$, of which node $i \in V$ is included in $\binom{N_v-1}{n-1}$, therefore the probability of node $i$ being in the sampled network is simply $\pi_i = \frac{n}{N_v}$, as in Equation 1. Since an edge $\{i,j\} \in E$ is then included in $E^*$ only if $i,j \in V^*$, the probability of edge $i,j$ being in the sampled network is $\pi_{\{i,j\}}$, as in Equation 1.

$$\pi_i = \frac{n}{N_v}, \qquad \pi_{\{i,j\}} = n(n-1)/N_v(N_v-1) \tag{1}$$

### Incident subgraph sampling

Incident subgraph sampling takes the opposite approach to induced subgraph sampling, in that it first selects $n$ edges from among the edge set $E$ and then observes the nodes incident with the selected edges, $\{i \in V : \{i, \cdot\} \in E^*\}$. The edges are again selected by simple random sampling without replacement, and so the probability of an edge from the population network $\{i,j\} \in E$ being included in the sampled network is derived in the same way as $\pi_i$ for induced subgraph sampling. The probability of an observed node $i \in V$, being included in the sampled network is slightly more complicated in this case, because it is included whenever one or more of its incident edges are included.

$$\pi_{\{i,j\}} = \frac{n}{N_e}, \qquad \begin{aligned} \pi_i &= \mathbb{P}(i \in V^*) \\ &= 1 - \mathbb{P}(\text{no edges incident to } i \text{ are sampled}) \\ &= \begin{cases} 1 - \frac{\binom{N_e - d_i}{n-1}}{\binom{N_e}{n}} & \text{if } d_i \leq N_e - n, \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

Where $d_i$ is the degree of node $i$. Note that when calculating node inclusion probabilities for incident subgraph sampling you must know the degree of the nodes in addition to the total number of

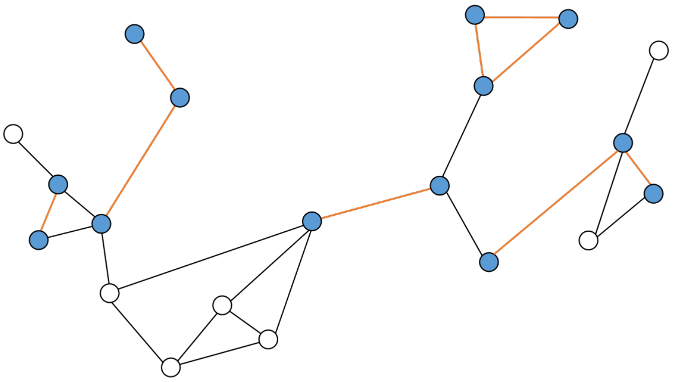

Figure 2: Incident subgraph sampling on a network of order 20. First the 9 arcs shown in orange are selected and then blue nodes are observed to form the sampled network.

edges in the network.

In a social network example, incident subgraph sampling can be thought of as selecting from a list of all connections, where connection is assumed here to be reflexive so that edges are undirected. Individuals are then observed if they are linked by any of the selected set of connections. In order to calculate node inclusion probabilities requires knowledge of the total number of connections involving each individual in addition to the overall total number of connections. Figure 2 illustrates incident subgraph sampling on a network of order 20.

### 3.2 Plug in Estimator

The aim of data collection on a network $G$ is usually to discover some feature of the network structure or network decoration, $\eta(G)$. Examples of the former may be the average degree, the total number of edges or node centrality, amongst others. Examples of decoration, on the other hand, may be the population proportion of each gender, or a measure of homophily in the network - how likely similar nodes are to be connected to one another.

Given a sampled network $G^*$, we can not hope to know $\eta(G)$ exactly, but we can attempt to estimate it by $\hat{\eta}(G^*)$. It is tempting to use a 'plug in' estimator of $\eta$, that is to use $\hat{\eta}(G^*) = \eta(G^*)$. This is an approach made implicitly when assuming that a network is representative of the larger population from which it is taken, but is shown to be problematic in the text by Kolaczyk [10]. Considering the average node degree as the characteristic of interest we have

$$\eta(G) = \frac{1}{N_V} \sum_{i \in V} d_i, \tag{3}$$

which we will use to demonstrate how using a plug in estimator might go awry. Suppose we have a sampled network of order $n$, so that we have a set of sampled vertices $V^* = \{i_1, \ldots, i_n\} \subset V$ and their *observed* degrees $\{d_i^* : i \in V^*\}$. The plug in estimator estimator of the average degree is then given simply by:

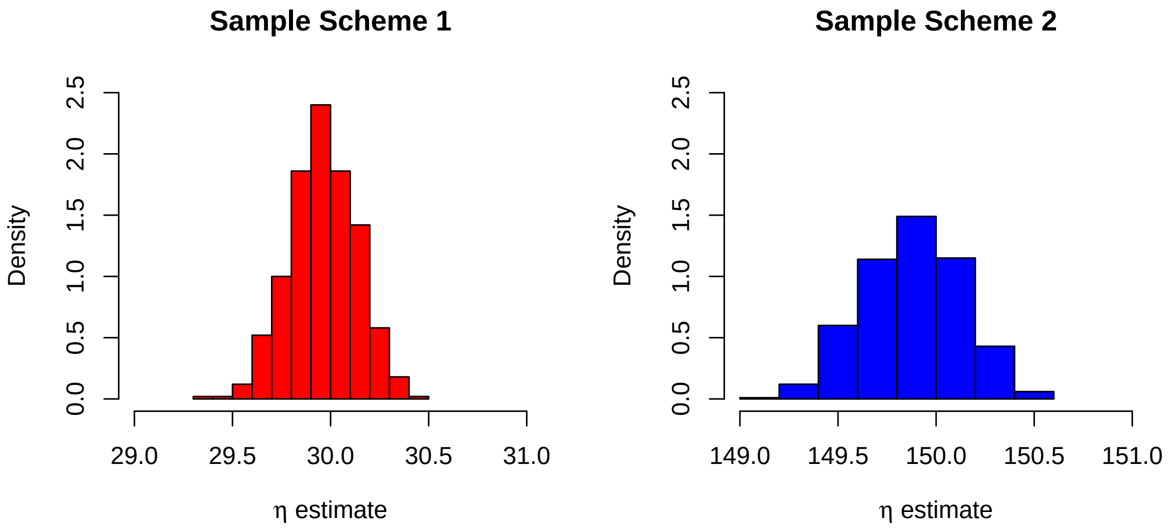

Figure 3: Histograms of mean degree estimates $\hat{\eta}$ over 500 sampled networks using two sampling schemes. Population mean degree $\eta(G) = 149.89$.

$$\hat{\eta} = \eta(G^*) = \frac{1}{n} \sum_{i \in V^*} d_i^*. \tag{4}$$

To demonstrate the pitfalls of the estimator, we consider its use under two induced subgraph sampling schemes. The first will be as described in Section 3.1, where a sampled node set $V_1^*$ is obtained through simple random sampling without replacement from the population network and the edge set is formed by the edges between these nodes, $E_1^* = \{\{i,j\} \in E : i,j \in V_1^*\}$. The second sampling scheme will obtain the sampled node set $V_2^*$ in the same manner but observe all edges incident with the sampled node set to be a part of the sampled edge set, $E_2^* = \{\{i,\cdot\} \in E : i \in V_2^*\}$.

While Kolaczyk looked at a protein interaction network for demonstration, we will consider the estimator when the population network is an Erdos-Renyi graph with 10,000 nodes and a probability of 0.015 for an edge forming between any pair of nodes. Figure 3 shows histograms of $\hat{\eta}$ for 500 sampled networks of 2,000 nodes from the population network by each of the sampling schemes. The estimate from the first sampling scheme underestimates the average degree by a factor of approximately the sampling proportion, which is because it only includes edges between selected nodes. The second sampling scheme is unbiased as we observe the true population degree of each node $d_i$, but this information is often unavailable when not all connections are part of a collective record.

This example illustrates two key points. Firstly, it shows the importance of the data collection mechanism when estimating properties of the population network. This can easily be forgotten, as so often problems begin with a data set without questioning the collection method. Secondly, this example motivates the many correction methods for the simple plug in estimator, in this case by multiplying by a factor of $\frac{N_v}{n}$. While for this sampling scheme correction is straightforward, the process can become complex for the sampling schemes used to collect data in practice.

## 3.3 Horvitz-Thompson Estimator

Here we make explicit the assumption made in Section 3.2 of design based inference. It is assumed that the measurements made on the sampled network are taken without error, though they may not capture the true state of the population network. This means that the only source of error in the experimental design comes from that introduced by the sampling scheme. An example of this error would be the bias introduced by the use of sampling scheme 1 with a plug in estimator of mean degree, as in Section 3.2 and shown in Figure 3.

The Horvitz-Thompson estimator of totals, first introduced in [9], provides one way to correct for sources of error caused by non-standard sampling designs. In considering an estimator of totals, we can consider more than the expected degree in a network. We could also look at, for example, the degree distribution of the network or the proportion of components (nodes or edges) with a given decoration.

### Theory

We begin in a general setting before going on to look at application to estimation of network properties in particular. Suppose that we have some population $\mathscr{U} = \{1,2,\ldots,N_u\}$ consisting of $N_u$ units, and that each of these units, $i \in \mathscr{U}$, has some property $y_i$ associated with it. Suppose also that a sample $S = \{i_1, i_2, \ldots, i_n\}$ of size $n$ has been taken from $\mathscr{U}$ according to some sampling scheme $\mathscr{S}$. A task prominent across statistics is then to estimate the population total of the property, $\tau = \sum_i y_i$, and by extension its average value or prevalence across the network, $\mu = {}^{\tau}\!/_{N_u}$.

The standard case of this problem would be to have $\mathscr{S}$ as simple random sampling with replacement, where the elements of $S$ are included from $\mathscr{U}$ with uniform probability. This leads to the natural unbiased estimators of population mean of $y$ by the sample mean, $\hat{\mu} = \bar{y} = ({}^{1}\!/_{n}) \sum_{i \in S} y_i$, and rescaling this achieves an unbiased estimator of population total $\hat{\tau} = N_u \bar{y}$. In practice, it is rare that the sampling scheme $\mathscr{S}$ which is intended and achieved will be simple random sampling with replacement. Amongst the many variations, the elements of $\mathscr{U}$ may have differing inclusion probabilities $\pi_i$, they may be sampled without replacement or both.

If it is the case that inclusion probabilities are unequal, then the sample mean will be a biased estimator of the population mean - the inclusion probability will not be equal to the population proportion. Horvitz-Thompson estimation is designed to counteract this by re-weighting the average to account for the differing inclusion probabilities. Letting $S$ now be the set of unique units sampled from $\mathscr{U}$, and $\pi_i$ be the inclusion probability of $i$ in $S$, the Horvitz Thompson estimate of $\tau$ is given by:

$$\hat{\tau}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i}. \tag{5}$$

This estimate can be easily shown to be unbiased, as in Equation 6. The estimate $\hat{\tau}_\pi$ leads to the unbiased estimator of the population mean $\hat{\mu}_\pi = \frac{\hat{\tau}_\pi}{N_v}$.

$$\mathbb{E}\left[\hat{\tau}_\pi\right] = \mathbb{E}\left[\sum_{i \in S} \frac{y_i}{\pi_i}\right] = \mathbb{E}\left[\sum_{i \in \mathscr{U}} \frac{y_i}{\pi_i} \mathbb{I}_{\{i \in S\}}\right] = \sum_{i \in \mathscr{U}} \frac{y_i}{\pi_i} \mathbb{E}\left[\mathbb{I}_{\{i \in S\}}\right] = \sum_{i \in \mathscr{U}} y_i = \tau. \tag{6}$$

It may otherwise be the case that the units of $\mathscr{U}$ are sampled uniformly at random without replacement. Then as we saw for induced subgraph sampling the inclusion probabilities of any single unit $i$ or pair of units $\{i,j\}$ from $\mathscr{U}$ in $S$ are given respectively by $\pi_i$ and $\pi_{\{i,j\}}$ as in Equation 1. This results in the same estimates of population total and mean as in the sampling with replacement case, $\hat{\tau}_\pi = N_v\bar{y}$ and $\hat{\mu}_\pi = \bar{y}$, though the estimates without replacement have smaller variance as every observation is distinct and contributes to lowering the uncertainty in the estimate.

### Application to degree estimation

Many network level properties may be viewed as a sum or average over the network by appropriate selection of the population units $\mathscr{U}$ and the unit property $y_i$. For instance, the average degree of the population graph, by nature of being a mean, is a scaled sum. Letting $\mathscr{U} = V$ and $y_i = d_i$ then $\bar{d} = (1/N_v)\sum_{i\in\mathscr{U}} y_i$. Now suppose that we have some sampled network $G^* = (V^*, E^*) \subseteq G$, sampled according to sampling scheme $\mathscr{S}$. Estimating the average degree in $G$ is then simple using a Horvitz-Thompson estimate, so long as we can calculate the inclusion probabilities of each element of $G$ being in $G^*$, $\pi_i$ for each node and $\pi_{\{i,j\}}$ for each edge.

Instead of considering only the average degree of a network we could instead look at the full degree distribution $f_d$. The degree distribution describes the proportion of nodes in the population network with any given number of incident edges. From our sampled network $G^*$ we see only the observed degree distribution, $f_d^*$. This can be corrected for by using the probability $\mathbb{P}(d^*,d)$ of a node with population degree $d$ being reduced to observed degree $d^*$ under the sampling scheme $\mathscr{S}$.

$$\mathbb{E}\left[f_d^*\right] = \sum_{d'=0}^{N_v-1} \mathbb{P}(d,d')f_{d'} \tag{7}$$

Equation 7 can then be used to determine a system of equations to solve for the population degree distribution $(f_0, f_1, \ldots, f_{N_v-1})$, whenever the number of sampled nodes exceeds the maximum degree in the population network [10]. In the case of where $\mathscr{S}$ is induced subgraph sampling Frank [6] showed the correction probabilities to be:

$$\mathbb{P}(d^*,d) = \frac{\binom{d}{d^*}\binom{N_v-1-d}{n-1-d^*}}{\binom{N_v-1}{n-1}}. \tag{8}$$

## 3.4 $\ell_2$ Risk Minimiser

In this section we explore an alternate estimator of node degree, derived from a risk theoretic perspective. The multivariate $\ell_2$ risk minimiser allows for the use of correlation in degree structure across the observed network to aid in estimation of true node degree. The $\ell_2$ risk minimiser was considered by Ganguly and Kolaczyk [7] for the estimation of individual node degree, as opposed to the average or distribution of degree over the network. The focus is on estimating degree, given that a node had been included in the sampled network through induced subgraph sampling (as described in Section 3.1). Comparison is drawn with the method of moments estimator of population degree.

**Frequentist Risk**

In a general estimation setting, the frequentist $\ell_2$ risk of an estimator $\hat{\theta}$ of a parameter with true value $\theta_0$ is defined as:

$$\mathcal{R}(\hat{\theta}, \theta_0) = \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right]. \tag{9}$$

The risk minimiser will therefore be the estimate which minimises the mean squared parameter estimation error.

### 3.4.1 Set-up & A Useful Result

We continue the use of $G^* = (V^*, E^*)$ to denote a sampled network of order $n$ from the population network $G = (V, E)$, which is of order $N_v$. Throughout this section the sampling scheme $\mathscr{S}$ used to sample $G^*$ from $G$ is induced subgraph sampling with known sampling proportion $p = \frac{n}{N_v}$. Relabelling the nodes of the population network so that $v_1, \ldots, v_n$ are included in the sampled network, we can denote the population network degree vector $(d_1, d_2, \ldots, d_{N_v})$ by $\mathbf{d}$ and that of the observed network, $(d_1, d_2, \ldots, d_n)$ by $\mathbf{d}^*$.

As we saw in Equation 1, the inclusion probability of each node in the population network under induced subgraph sampling is $\pi_i = \frac{n}{N_v} = p$ and so the observed degree of node $i$, $d_i^*$ follows a $\text{Binomial}(d_i, p)$ distribution, as each of the $d_i$ edges is included if and only if the other node it is incident to is also sampled. This leads to the method of moments estimator $\hat{d}^{MME} = \frac{d_i^*}{p}$, which was mentioned in Section 3.2. The distribution of observed degree also motivates Results 10 and 11, which will be of use in the derivation of the multivariate risk minimiser. For a proof of the results see the appendix of Ganguly and Kolaczyk [7].

*Under induced subgraph sampling, the mean and covariance matrix of the observed degree vector are*

$$\mathbb{E}[\mathbf{d}^*] = p\mathbf{d}, \tag{10}$$

$$\text{Var}[\mathbf{d}^*] = p(1-p)D, \tag{11}$$

*where $D$ is an $n$ by $n$ matrix with $D_{ij} = d_i$ if $i = j$, otherwise $D_{ij}$ is the number of common neighbours of nodes $i$ and $j$ in $G$.*

### 3.4.2 Univariate Risk Minimisation

We first look at estimating the degree of each node individually, by minimising the associated univariate risk using a scale up estimator of the form $\hat{d}_i = c_i d_i^*$, for some scalar $c_i$ which is to be determined. Since we know that the $d_i^* \sim \text{Binomial}(d_i, p)$ and we are considering $\ell_2$ risk, it follows that:

$$\mathcal{R}(\hat{d}_i, d_i) = \text{Bias}^2(\hat{d}_i) + \text{Var}(\hat{d}_i) = \text{Bias}^2(c_i d_i^*) + \text{Var}(c_i d_i^*) = (c_i p d_i - d_i)^2 + p(1-p)c_i^2 d_i. \tag{12}$$

Differentiating the expression for univariate risk in Equation 12 with respect to $c_i$ and equating to zero, we find the risk minimising estimator of $c_i$:

$$\hat{c}_i = \frac{d_i}{pd_i + 1 - p}$$

which can be shown by differentiating again to minimise the $\ell_2$ risk whenever $d_i > \frac{1-p}{p}$. Plugging the method of moments estimator of $d_i$ into this expression we get the univariate risk minimising estimator of $d_i$:

$$\hat{d}_i = \frac{d_i^{*2}}{p(d_i^* + 1 - p)}. \tag{13}$$

By taking the Taylor expansion of the formula for $\hat{d}_i$, we can see that the estimator is biased. The Taylor expansion is valid for all non-zero observed degrees, and so is applicable when the observed degree is concentrated about its mean, a condition which relaxes as the average observed degree increases.

$$\mathbb{E}\left[\frac{d_i^{*2}}{p(d_i^* + 1 - p)}\right] = \frac{1}{p}\mathbb{E}\left[d_i^*\left(1 + \frac{1-p}{d_i^*}\right)^{-1}\right] \approx \frac{1}{p}\mathbb{E}\left[d_i^*\left(1 - \frac{1-p}{d_i^*}\right)\right] = d_i - \frac{1-p}{p}$$

Making a bias correction on the estimator $\hat{d}_i$ leads to an estimator $\hat{d}_{i,\mathrm{u}}$ which has comparable risk to the method of moments estimator when the population degree is small but has lower risk when the population degree exceeds $\frac{1-p}{p}$ [7]. The bias corrected estimator is then given by:

$$\hat{d}_{i,\mathrm{u}} = \frac{d_i^{*2}}{p(d_i^* + 1 - p)} + \frac{1-p}{p}. \tag{14}$$

### 3.4.3 Multivariate Risk Minimisation

Rather than minimising the $\ell_2$ risk associated with the estimation of each node's degree individually, we can consider minimising the risk over the full set of degrees to be estimated. This follows as an extension to the univariate unbiased risk minimiser derivation in the previous section, but allows for exploitation of the covariance structure between observed degrees from Result 11 during the estimation procedure.

We now consider minimising the sum of the $\ell_2$ risk across the set of sampled nodes using an estimator of the form $\hat{\mathbf{d}} = A\mathbf{d}^*$, where $A$ is an $n$ by $n$ matrix to be determined. Again, because we are using the $\ell_2$ risk, the risk associate with the estimator is equal to the sum of the squared bias of the estimator and its variance. Using Results 10 and 11, we have that:

$$\begin{aligned}\mathcal{R}(\hat{\mathbf{d}}, \mathbf{d}) &= \mathrm{Bias}^2(A\mathbf{d}^*) + \mathrm{Var}(A\mathbf{d}^*) \\ &= (pA - I_n)\mathbf{d}\mathbf{d}^T(pA - I_n) + p(1-p)ADA^T \\ &= A\left(p^2\mathbf{d}\mathbf{d}^T A^T + p(1-p)D\right)A^T - p(\mathbf{d}\mathbf{d}^T A^T + A(\mathbf{d}\mathbf{d}^T) + \mathrm{constant}.\end{aligned} \tag{15}$$

The multivariate risk minimiser is then given by:

$$\hat{A} = \arg\min_{A} \sum_{i=1}^{n} \mathbb{E}(\hat{d}_i - d_i)^2 = \arg\min_{A} \ \text{trace}\left(\mathcal{R}(\hat{\mathbf{d}}, \mathbf{d})\right).$$

To find the value of $\hat{A}$, differentiate the trace of the estimator's risk and equate to zero. This yields:

$$\hat{A} = p\mathbf{d}\mathbf{d}^T \left(p^2\mathbf{d}\mathbf{d}^T + p(1-p)D\right)^{-1}. \tag{16}$$

Of course, we do not know the true values of **d** or $D$, and so we substitute for their method of moments estimators into Equation 16. Doing so gives the multivariate risk minimising estimator of the the degree vector as

$$\hat{\mathbf{d}}_m = \frac{1}{p}\mathbf{d}^*\mathbf{d}^{*T} \left(\mathbf{d}^*\mathbf{d}^{*T} + D^*\right)^{-1} \mathbf{d}^*. \tag{17}$$

Here $D^*$ is an $n$ by $n$ matrix with $D^*_{i,j} = d^*_i$ when $i = j$, otherwise $D^*_{ij}$ is the number of common neighbours of nodes $i$ and $j$ in the sampled network $G^*$. This estimator of the degree vector shrinks the method of moments estimator $\hat{\mathbf{d}}_{MME} = \frac{1}{p}\mathbf{d}^*$, and has been shown to outperform the method of moments estimator in expectation when the sampled graph is sparse, but with each sampled node having degree at least one [7].

### 3.4.4 Application to Common Network Structures

In addition to proving theoretically the conditions for risk minimising estimators outperform the method of moments estimators, Ganguly and Kolaczyk [7] consider their application to two common network models and an example network. The two network models considered are the Erdos-Renyi and Scale Free models, as described in Section 2. The example network is of human trafficking, made up of 31,428 nodes of which 12,387 were flagged as being suspicious and the aim was to estimate the degree of these flagged nodes.

For each of the population network structures, 50 population networks were simulated with each combination of four edge densities and sampling proportions, and then these population networks were sampled from using induced subgraph sampling. For each combination of density and sampling proportion, the method of moments, univariate and multivariate risk minimising estimators were compared for their average $\ell_2$ distance from the known true degree vectors over the 50 sampled networks. For the human trafficking network, it was assumed that the full network, which was generated by a Memex search, was unknown and sampled from this network 50 times to produce 50 smaller 'population' networks on which to test each of estimators at varying sampling proportions.

The results of the simulation study showed that the method of moments estimator was outperformed, on average, for all combinations of edge density and sampling proportions. The multivariate risk minimiser performed the best on the Erdos-Renyi simulated networks, while the univariate risk minimiser performed the best on the majority of cases for the Scale Free simulated networks and all sampling proportions of the example network. This gives weight to the proposed estimators over the method of moments estimator, it appears that the assumptions required for the risk minimising estimators to outperform the method of moments do indeed hold in example networks of interest.

### 3.4.5 Criticisms

It has been shown that in the case of the human trafficking network that the risk minimising estimators of the degree vector outperform the method of moments estimator. However, the conditions required for this to be the case point to some limitations in the applicability of the estimators. Presumably, the network is being sampled from using induced subgraph sampling because it is either too financially or computationally expensive to handle the entire network. The first case requires care during implementation but the latter brings up two issues.

Firstly, it is required that each of the sampled nodes be of at least observed degree one. This is unlikely to be the case when taking a small sampling proportion of a large network, for example from an online social network such as Facebook with 1.8 billion users [1]. Even in a sample as large as 100,000 users the assumption of on isolated nodes hold with prohibitively small possibility. This is because of the global sparsity of the network, with population average degree of around 250 (Ugander et al. [12]) and an upper limit of 5000 on degree. Secondly, the univariate risk minimiser would only outperform that of the method of moments when the observed degree of a node exceeded 18,000 connections.

## 4 Hidden Network Methods

In Section 3, we considered situations where the node set $V$ or the edge set $E$ of the population network was available to be sampled from, but where the size of the population network made it prohibitively expensive in terms of finance or computational power to survey the full network. In this section, we will consider circumstances where a sampled network is used because the population network is not available in its entirety.

Examples of hidden networks occur whenever there is not a readily available sampling frame for the population of interest. Criminal and terrorist networks provide good examples which are deliberately hidden, however it is not only surreptitious individuals who form hidden networks. The socially deprived, marginalised and elite all form groups who are not easily sampled from. Atkinson and Flint [2] suggest such groups as; young out of work men, people with a rare or stigmatised illness and the homeless as examples of hidden networks. For various reasons these groups are difficult to contact or survey, usually because they are unknown to the researcher or unwilling to participate in studies.

Section 4.1 explores two sampling schemes which could be used to reach hidden populations, namely snowball sampling and respondent driven sampling. Section 4.2 then goes on to describe how issues caused by respondent driven sampling might be addressed through bootstrap techniques.

## 4.1 Sampling schemes for hidden networks

In order to reach hidden populations, non-conventional sampling schemes are often necessary. These schemes frequently involve some referral referral mechanism, which is initiated from one or more initial locations in the hidden network.

### Snowball sampling

To initiate snowball sampling, $n$ nodes from the hidden network are identified and included in the initial sampled node set $V_0^*$. Next, all edges incident to a sampled node are observed and included in the initial sampled edge set $E_0^*$. The sampled node set $V_0^*$ is then augmented with the nodes incident to $E_0^*$, to give $V_1^* \supseteq V_0^*$. This completes the first wave of sampling, giving the sampled network $G_1^* = (V_1^*, E_0^*)$.

Subsequent waves of sampling proceed in the natural recursive manner, with the $k^{\text{th}}$ wave of sampling proceeding as follows. Include the edges which are incident to the newly added nodes $V_{k-1}^* \setminus V_{k-2}^*$ to update the edge set giving $E_{k-1}$, then update the node set to $V_{k+1}^*$ which includes the nodes incident to the newly included edges $E_{k-1} \setminus E_{k-2}$. In practice, waves may continue until the sampling 'dies out' when no new nodes or edges are added, or else a set number of waves may be performed. Figure 4 demonstrates two wave snowball sampling on a network of order 20.

In the case of only one wave of sampling taking place, snowball sampling is often referred to as star sampling. It is in this case that the calculation of node and edge inclusion probabilities is simplest, but even then it is only possible under the assumption that the inclusion probabilities of the $n$ initiating nodes are known. The initiating nodes are often assumed to be selected using simple random sampling without replacement or by a sequence of Bernoulli random trails, which are approximately equivalent when the number of nodes is large and the sampling probability is small [10]. However, both of these assumptions are usually unfounded, as there is no reason to believe that all members of the network would be equally available to the researcher. The assumption is usually being made for the convenience in calculating inclusion probabilities so that Horvitz-Thompson or some other form of estimation may be used.

Snowball sampling is the most common sampling scheme used to access hidden populations, with the initiation nodes being selected by ease of availability rather than being draw randomly from the hidden network. For this reason snowball sampling suffers heavily from bias. Firstly, there is the bias from the initiating group being by definition 'more cooperative' or 'less hidden', and therefore being unrepresentative of the population network. Secondly, there is an unknown degree of bias induced in the referrals by individuals who are 'protecting' one another through not connections. This is of particular advantage in persecuted networks where individuals face being further stigmatised for 'snitching'. Finally, there is a bias toward the nodes of higher degree, as they are more likely to be referred and that nodes of lower degree are more likely to be excluded. These biases stemming from the sampling design are not easily corrected for, and so snowball sampling is considered to be a form of convenience sampling for network structures, producing biased and inconsistent samples [8].

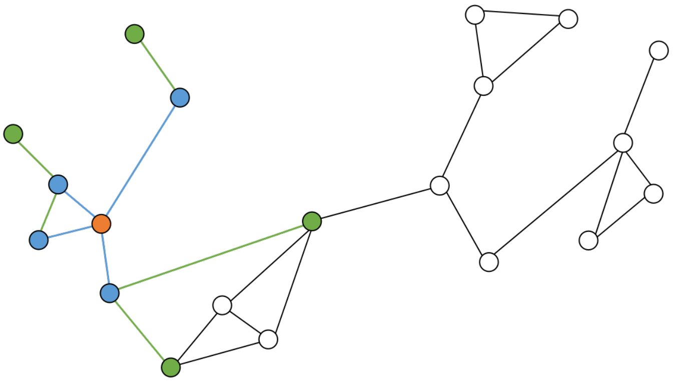

Figure 4: Two wave snowball sampling on a network of order 20. The single initiation node is shown in orange, with nodes and edges sampled in the first wave shown in blue, and those sampled in the second wave shown in green.

### 4.1.1 Respondent driven sampling

Respondent driven sampling is an alternate form of referral sampling on hidden networks, first developed to target injective drug users as part of an AIDS intervention effort by Broadhead and Heckathorn [5]. Again, respondent driven sampling relies on a well connected population network to refer individuals into the sample, but circumvents the issue of dependency on the initialisation of the sampling. Samples obtained by respondent driven sampling are independent of where the sample begins, they reduce the bias associated with initail volunteering, and provide a means to control for individual node degree [8].

To initiate a respondent driven sample, $n$ individuals from the hidden network are identified and recruited to the sampled node set $V_0^*$. These individuals are then given an incentive to recruit others from the hidden network by giving them uniquely numbered recruitment tokens. The incentives for recruiting may be tangible, such as a cash reward, or intangible, such as greater community protection from disease. Any individuals recruited will then be linked to their referrer, the edge linkging them will be added to $E_0^*$ and the new recruit will be given the chance to recruit others themselves. This is then repeated through $k$ generations of recruitment until the sampling 'dies out' or until the required sample size is reached. This results in the sampled network $G^* = (V_k^*, E_k^*)$ with a tree or forest structure, depending on the number of initialising nodes.

This method of sampling has the benefit that the dependence on the initial nodes is weakened by the iterative recruitment process. Also the issue of censoring for protective purposes is reduced, as respondents are not asked directly for a list of connections [4]. Figure 5 shows respondent driven sampling on a hidden network of order 20.

## 4.2 Tree bootstrap for respondent driven sampling

Respondent driven sampling partially resolves many of the issues inherent in a snowball sampling scheme. We now look at how it may be used to estimate the hidden network average or prevalence of some node decoration.

Under five conditions on the population network, respondent driven sampling may be con-

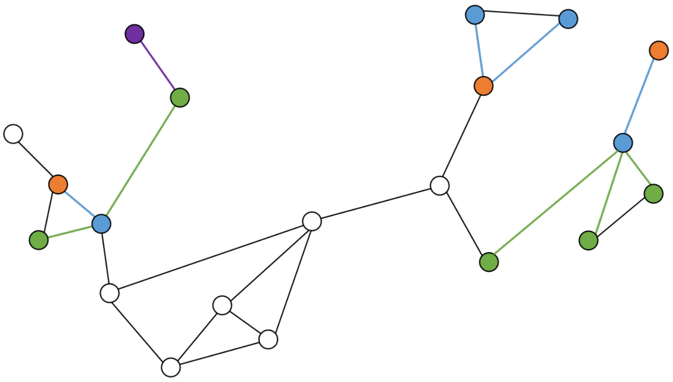

Figure 5: Respondent driven sampling on a network of order 20. Initiation nodes are shown in orange, the nodes and edges added to $V^*$ in the following successful recruitment in rounds 2, 3, and 4 are shown in blue, green and purple respectively.

sidered as a first-order Markov chain on the population network. These conditions are; the population network must be finite order, undirected and connected; the nodes from the population network may be resampled; recruitment is uniform on the neighbours of the recruiter and the population degree of each sampled node is accurately recorded. The stationary distribution of this chain is then proportional to the population degree of the nodes [4]. This motivates the estimation of some node decoration $x_i$ using the Volz-Heckathorn estimator which, like a Horvitz-Thompson estimator, reweights observations because of unequal sampling probabilities. The Volz-Heckathorn estimator of population mean or prevalence of node decoration $\mu_x$ is given in in Equation 18.

$$\hat{\mu}_{VH} = \left( \sum_{i=1}^{N_v} \frac{x_i}{d_i} \right) \left( \sum_{i=1}^{N_v} \frac{1}{d_i} \right)^{-1} \tag{18}$$

Due to the dependence between samples in respondent driven sampling, the variability of this estimator and therefore calculating its confidence intervals is not straightforward, unlike for induced or incident subgraph sampling. This is usually worked around by using bootstrapping techniques to resample the observed data, creating pseudo data sets over which the variance of the estimator can itself be estimated. In doing so, many methods try to exploit the Markov structure of the sampling scheme, assuming that this holds when the data are aggregated by decoration value. This property is not ensured theoretically and often leads to gross underestimates of the variability of the estimator, leading to 95% confidence intervals with true coverage rates as low as 40% and rarely exceeding 70%. This is demonstrated by Baraff et al. [4], who propose a tree bootstrap method for estimating the variance of the estimator which goes some way to mitigating this issue.

### Tree Bootstrap

Tree bootstrapping is a hierarchical bootstrapping method. In the first bootstrap stage, the initial seeds are resampled with replacement to produce a bootstrap seeding. From there the recruits of each of these seeds are resampled with replacement to fill in the next generation, giving a bootstrap first generation.

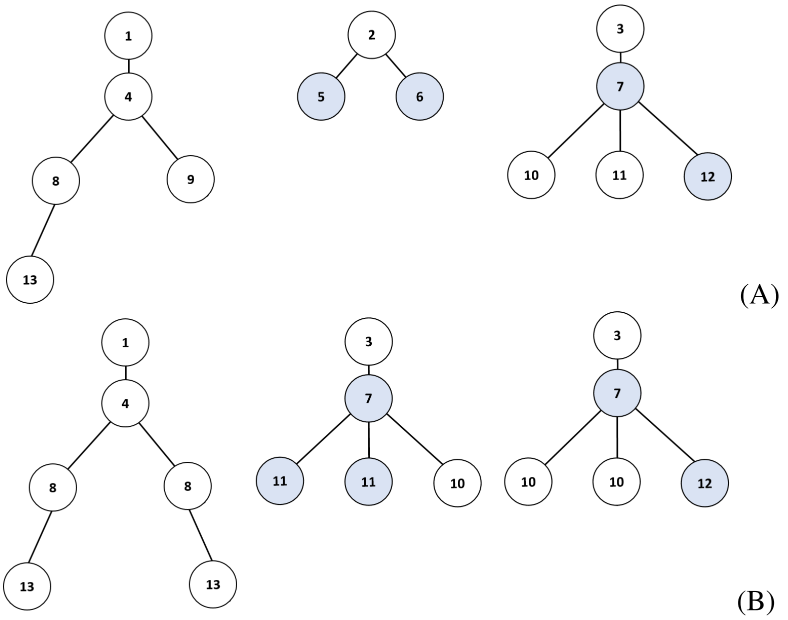

Figure 6: (A) Respondent driven sample from Figure 5 with personal identifiers and decorated with sex of respondents shown. (B) One tree bootstrap resampling of the recruitment trees.

The recruits of each of the bootstrap first generation are then resampled to form the bootstrap second generation and so on until no further recruits are available. This is demonstrated in Figure 6 using the respondent driven sample trees from Figure 5. In Figure 6, the individuals are coloured according to their decoration value (e.g. sex) rather than the order in which they are sampled; this is shown in the tree structure from the top of the diagram downward. From many such bootstrapped trees the variance of any network statistic of interest may be estimated.

### Results of simulation study

Baraff et al. [4] compared the tree bootstrap method to multiple existing methods of variance estimation. This was done using simulated respondent driven sampling generated from several large and completely enumerated networks. This was done so that the performance of the variance estimators could be compared with one another, using the true coverage of their resulting confidence intervals. The bootstrap intervals were shown to be clearly superior, always achieving true coverage closer to the stated 95% than any of the other estimators. The improvement over the other estimators was significant, as in the majority cases others failed to reach even 70% true coverage, where as the tree bootstrap estimator achieved an average of around 80% coverage. While in some cases the confidence intervals from the bootstrap estimator were too wide to be useful in practice, this is indicative of high dependence between the observations and a low effective sample size. In these cases it is the respondent driven sampling data itself which is unsuitable, and not the estimation methods.

## 5 Review and Further Work

As we have seen, partially observed networks arise in a variety of current research areas. It is frequently the case that a partially observed network is assumed to be representative of the population network it is representing. It was shown in Section 3.2 that this assumption can cause issues of severe bias when extrapolating the partially observed network properties to the population network, unless the sampling scheme used to generate the partially observed network is taken into consideration.

For networks with a readily available sampling frame, but which are too large to be surveyed entirely we explored how estimation of population properties could be estimated following induced- and incident subgraph sampling. In Section 3 this was done in two ways, firstly using Horvitz-Thompson estimators and the secondly by minimising the $\ell_2$ risk associated with the estimator. Concerns were raised about the applicability of the risk minimising estimators, due to the conditions required to perform better than the method of moments estimator. It would be an interesting area of further research to establish the conditions in which the estimator is favourable under sampling schemes other than incident subgraph sampling, to see if these would prove to be less restrictive.

For networks with no available sampling frame, we considered in Section 4 the methods of snowball and respondent driven sampling for drawing a sample from the hidden network. The former of these sampling schemes suffers heavily from biases, but the latter allows for estimation of population network properties. The variability of these respondent driven sample estimates are often underestimated, an issue which tree bootstrapping was seen to overcome to some extent. The theoretical properties of the tree bootstrap estimator are not well explored, presenting a potential area of expansion in the literature. IN addition, exploring the application of bootstrap techniques to other sampling designs could be a productive area of research.

The literature on networks is spread through the journals of many fields because of their wide applicability. These fields include but are not limited to; mathematics, statistics, operations research, biology and economics. The respective approaches of these fields to network analysis are quite disparate, as they each approach the topic with their own perspective and expertise. It may be of great benefit to all fields to compile an interdisciplinary review, outlining and signposting the areas which have been well explored in each field, and current open questions of interest. This would perhaps encourage more cross-disciplinary work and inform a more unified approach to future research across all interested fields.

## 6 References

[1] Facebook Newsroom statistics. `https://newsroom.fb.com/company-info/`. Accessed: 04/04/2017.

[2] Rowland Atkinson and John Flint. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, 33(1):1–4, 2001.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999.

[4] Aaron J Baraff, Tyler H McCormick, and Adrian E Raftery. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proceedings of the National Academy of Sciences*, page 201617258, 2016.

[5] Robert S Broadhead and Douglas D Heckathorn. Aids prevention outreach among injection drug users: Agency problems and new approaches. *Social Problems*, 41(3):473–495, 1994.

[6] O. Frank. *Statistical Inference in graphs*. PhD dissertation. Stockholm University, 1971.

[7] Apratim Ganguly and Eric Kolaczyk. Estimation of vertex degrees in a sampled network. *arXiv preprint arXiv:1701.07203*, 2017.

[8] Douglas D Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 44(2):174–199, 1997.

[9] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[10] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387881461. URL `https://books.google.co.uk/books?id=Q-GNLsqq7QwC`.

[11] Michael Salter-Townshend, Arthur White, Isabella Gollini, and Thomas Brendan Murphy. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4):243–264, 2012.

[12] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.