LSSP-2018-0450 Minor revision
Horvitz-Thompson estimator under partial information with an application to network degree distribution*

Alejandro Izaguirre† Gabriel Montes-Rojas‡

9th November 2018

Abstract

We present an extension of the Horvitz-Thompson estimator for estimating the total number of sets of a given size within a population. We study the limitations of the Horvitz-Thompson under partial information where we have a sample of population elements rather than of sets. The developed estimator is the chained Horvitz-Thompson. We apply this estimator to the estimate the network degree distribution to be used under probability sampling designs, in particular, induced subgraph and incident subgraph. Finally, we present Monte Carlo simulations to assess the accuracy of the proposed estimator.

Key words: *Horvitz-Thompson estimator, networks, network sampling designs, degree distribution, average degree.*

JEL classification: C13, C4.

*The authors are grateful to the Editor in Chief Prof. Narayanaswamy Balakrishnan and to an anonymous referee for suggestions to improve this paper.

†PhD student at Universidad de San Andrés, Buenos Aires, Argentina, e-mail: izaguirre.ale@gmail.com

‡Corresponding author: Instituto Interdisciplinario de Economía Política de Buenos Aires (IIEP-BAIRES-UBA) and CONICET, Facultad de Ciencias Económicas, Universidad de Buenos Aires, Córdoba 2122, 2do piso, C1120AAQ, Ciudad Autónoma de Buenos Aires, Argentina, e-mail: gabriel.montes@fce.edu.ar

1 Introduction

A growing literature related to social networks and their implications in economic outcomes emerged during the last years (see Jackson, 2008). A network represents a set of connections (edges) among a collection of agents (nodes). Most networks investigated today are parts of much larger networks. Although many applied works speak of *the network* when presenting empirical results, frequently it is only a sampled version of some larger underlying network. Sampling is of particular interest in the context of online social networks because they are usually very large.

Although the first works on network sampling started in the late 1960s by the hand of Ove Frank and his colleagues, the subject showed a growing interest in the last years in a number of fields such as economics, epidemiology, statistics, sociology and computer science, among others.

There are many papers in which the focus is on understanding the extent to which characteristics of a sampled network correspond to the complete network (see Zhang et al., 2015, for a literature review). Focusing on the analysis of the internet topology, this issue is studied in Lakhina et al. (2003) and Achlioptas et al. (2009). Typical characteristics of interest include degree distribution, density, diameter, clustering coefficient, average path length, among others. In general, under many sampling schemes, these measures are biased when we use sampled networks. One of the first works in proposing an estimator for network characteristics (average degree and density) based on a sampled network was Granovetter (1976).

Network models are widely used to represent information among interacting units and the structural implication of these relations, and the estimation of such models based on sampled networks suffers bias problems too. See, for instance, Handcock & Gile (2010), Santos & Barrett (2008) and Chandrasekhar & Lewis (2011). In the latter, the authors not only show how the bias arise when we use sampled data in the estimation of econometrics models, but also propose procedures to correct such biases. One of them is based on a graphical reconstruction, that is, using sample information to predict the full network. This approach requires a series of assumptions about the network formation process.

In this paper we focus our attention on the network degree distribution. The degree distribution of a network is a description of the relative frequencies of nodes that have different degrees. This measure is one of the most fundamental characteristic associated with a network and it is affected

by sampling, sometimes dramatically.

Network sampling is a highly relevant topic in the field of network science. Some references for the sampling literature are Rothenberg (1995), Handcock & Gile (2010), (Kolaczyk, 2009, ch.5), Ahmed et al. (2014), among many others. In this paper we restrict our attention to probability sampling designs.

The main problem we face in this article is given by the lack of information. Briefly, we are interested in sets but we do not observe sets, we observe elements of those sets instead. Given this issue, in the first part we extend the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952) to make it feasible under this setting. In the second part, we propose an estimator for the network degree distribution to be used under any probability sampling designs, and then we apply it for two widely used probability sampling designs known as induced subgraph (i.e., nodes sampling) and incident subgraph (i.e., edges sampling). A related work is Frank (1977), which presents an adaptation of the HT estimator to be used in networks contexts, but under the same framework as the original HT estimator. Finally we do a Monte Carlo simulation to assess the analytical results.

The paper is organized as follows. Section 2 presents the limitations of the Horvitz-Thompson estimator under partial information and Section 3 the corrected estimators. Section 4 develops the chained Horvitz-Thompson estimator. Section 5 applies the proposed estimators to the network degree distribution. Section 6 presents Monte Carlo simulations. Section 7 concludes.

2 The Horvitz-Thompson estimator under partial information

The HT estimator is a well known estimator mostly used for estimating population totals under probability sampling designs (see Fuller, 2011). In this paper we are interested in estimating population totals, but the HT estimator is unfeasible under the framework we work with.

The elements of the main population are grouped into sets and we want to know how many sets of a given size there are into the population. The issue for using the HT estimator is because the sample we have has only partial information about the sets. In particular, we have a sample of elements of the population, which means that we do not have a sample of sets, rather a sample of elements of the sets.

Consider a population and a collection of sets given by the following definitions.

Definition 1. Let $U = \{u_1, u_2, \ldots, u_N\}$ be a population of size $N < \infty$ whose elements are grouped into $J < \infty$ sets (not necessarily exclusive).

Definition 2. Let $\Theta = \left\{\theta^1, \ldots, \theta^J\right\}$ be a collection of the J sets in which the elements of U are grouped, with generic element $\theta = \{u_j, u_h, \ldots, u_l\}$, $1 \leq j, h, \ldots, l \leq N$, θ_k is a generic element of size k, with $k = 1, 2, \ldots, p$ and T_k is the total number of sets of size k in Θ.

Θ contains all the sets in which the elements of U are grouped into, such elements are not necessarily exclusive.

The total number of sets of size k in Θ can be written as $T_k = \sum_{\theta \in \Theta} I\left[\#\theta = k\right]$, where $\#\theta$ denotes the cardinality of θ, that is, the number of elements in θ. We want to estimate how many sets of size k there are into the population, that is, we want to estimate T_k. If we had a sample of Θ we could use the HT estimator as follows,

$$\hat{T}_k^{HT} = \sum_{\theta \in \Theta^s} \pi_\theta^{-1} I\left[\#\theta = k\right], \tag{1}$$

where $\Theta^s \subset \Theta$ is a sample of Θ and π_θ is the probability of selection of θ. If we have a sample of U instead of Θ the HT estimator is unfeasible because we do not know θ.

In this paper we propose an unbiased estimator for T_k (and its variance) based on the HT estimator to be used when we have a sample of U instead of Θ. The only requirement is that the sample has certain information about θ (see Assumption 1).

As an example suppose we want to know the total number of blocks in a city with different numbers of houses, that is, how many blocks are with, for example, 20 houses, how many with 25, and so on. Under this setting our population U is given by the N houses of the city which are grouped into J blocks, being θ a generic block and θ_k a generic block with k houses. Θ is a set (or collection) that contains all θ, and T_k is the total number of blocks with k houses. If we have a probability sample of Θ (a sample of blocks) we could use (1) for estimating T_k, but suppose we have instead a probability sample of U (a sample of houses instead of a sample of blocks), if we do not observe θ the estimator (1) is unfeasible.

3 Extending the HT estimator for partial information

We have a population whose elements are grouped into sets, and our main goal here is to estimate how many sets of a given size there are in the population. The problem is that we partially observe those sets, that is, we only observe subsets.

For an intuitive introduction to the methodology proposed here suppose we divide each set of the population Θ into all possible combinations, that is, we generate all possible subsets from *each set* of Θ, then suppose we arrange these subsets into a new population named Γ. In other words, we create a new population given by all subsets we can get from *each set* of Θ.

The total number of subsets of a given size in Γ is a linear combination of the total number of sets in Θ. This relation is the key for the estimator we propose here. Our proposal consists in estimating the total number of subsets in Γ (since we observe them by assumptions), and then to estimate the total number of sets in Θ based on the implied linear relationship.

Let L_c be the total number of subsets of size c in Γ, we have that $L_c = \sum_{k=1}^{p} T_k \binom{k}{c}$. As noted above, the total number of subsets of size c in Γ, that is L_c, is a linear combination of the total number of sets in Θ, T_k.

As an example, suppose Θ contains 4 sets of size 5 and 3 sets of size 4, that is, $T_5 = 4$ and $T_4 = 3$, so there are $L_4 = 4\binom{5}{4} + 3\binom{4}{4}$ subsets of size $c = 4$ in Γ.

Generalizing the previous statement we have the following equality

$$L = AT, \tag{2}$$

where $L = (L_1, L_2, \ldots, L_p)'$ is a $(p \times 1)$ vector, $A = [a_{ij}]$, where $a_{ij} = \binom{j}{i}$ with $i, j = 1, \ldots, p$ is a $(p \times p)$ upper unitriangular matrix (the matrix A is a submatrix of the Pascal Matrix, we present some properties about A in the Appendix), and $T = (T_1, T_2, \ldots, T_p)'$ is a $(p \times 1)$ vector.

The element T_k of T is the total number of sets of size k in Θ, the element L_c of L is the total number of subsets of size c in Γ. The matrix A transforms the total number of sets in Θ into the total number of subsets in Γ. The previous equation summarizes the linear relation between the total number of sets (of different sizes) and the total number of subsets (of different sizes) we can

get from them.

Our methodology consists in estimating L and then replace it in $T = A^{-1}L$ to get an estimator for T. Given that A is a non stochastic upper unitriangular matrix its inverse exists, furthermore, $A^{-1} = \left[\bar{a}_{ij}^{-1}\right]$ where $\bar{a}_{ij}^{-1} = (-1)^{j+i} \begin{pmatrix} j \\ i \end{pmatrix}$.

3.1 The Horvitz-Thompson estimator for L_c

Consider the following definitions and assumptions.

Definition 3. Let S be a probability sample of U. Let $\mathcal{S} = 2^U$ be the σ-field of all possible samples of U, and Π be a probability measure, $\Pi : \mathcal{S} \mapsto [0, 1]$. The triplet $(U, \mathcal{S}, \Pi)$ is a probability space.

Our raw material here are the subsets we can get from the sets of Θ, we construct an "artificial" population of such subsets as follows:

Definition 4. Let $\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$ be a collection of all possible subsets we can get from ***each element*** of Θ where $P(\theta)$ is the power set of θ, γ is a generic element of Γ, and γ_c a generic element of size c, with $c = 1, 2, \ldots, p$. The total number of elements of size c in Γ is given by $L_c = \sum_{\gamma \in \Gamma} I[\#\gamma = c]$.

Although the elements of Γ (and Γ^s) are sets we call them subsets to highlight that they are subsets of θ. To avoid confusions we refer to the elements of Θ as sets and to the elements of Γ (and Γ^s) as subsets.

Definition 5. Let $\Gamma^s = \{\gamma | \forall \gamma \in S\}$ be the collection of all subsets γ in sample S, $\Gamma^s \subseteq \Gamma$. Let $p^s \leq p$ be the maximum size of $\gamma \in S$.

To clarify, S is a sample of U, and $\gamma \subseteq \theta$ are subsets composed by elements of U. What we mean by $\gamma \in S$ is that all $u \subseteq \gamma$ are in S.

Definition 6. Let $\pi_\gamma = Prob(\gamma \in \Gamma^s)$ be the probability of selection of γ and $\pi_{\gamma\gamma'} = Prob(\gamma \cup \gamma' \in \Gamma^s)$ be the joint probability of selection of γ and γ'. Given definition 5, $\pi_\gamma = Prob(\gamma \in S)$ and $\pi_{\gamma\gamma'} = Prob(\gamma \cup \gamma' \in S)$.

Consider now the following assumptions.

Assumption 1. *For all $u \in S$ we can identify every θ in Θ such that $u \in \theta$.*

Assumption 2. $\pi_\gamma > 0 \; \forall \, [\#\gamma \leq p] \in \Gamma$. *The probability of selection of any element in Γ has to be positive.*

Assumption 3. $\pi_{\gamma\gamma'} > 0 \; \forall \, [\#\gamma \leq p] \vee [\#\gamma' \leq p] \in \Gamma$. *The joint probability of selection of any pair of elements in Γ has to be positive.*

Assumption 1 is the key assumption for the proposed estimator. It does not imply that we know all the elements in θ, it only assumes we know to which set (or sets) each element of the sample belongs. In other words, we can match every $u \in S$ with any $\theta \in \Theta$ such that $u \in \theta$. The idea behind this assumption is that we can group the elements in S as they are grouped in Θ, and thus we have a sample of subsets of θ. These subsets could indeed be θ, but we do not know it a priori.

Consider the following example. Let $U = \{u_1, u_2, u_3, u_4, u_5\}$ be a population of individual elements, let $\Theta = \left\{\theta^1, \theta^2, \theta^3\right\}$ be a population of sets, where $\theta^1 = (u_1, u_3)$, $\theta^2 = (u_2, u_3, u_5)$, $\theta^3 = (u_4, u_5)$, and let $S = \{u_1, u_2, u_5\}$ be a sample of U. Assumption 1 assumes we know that u_1 belongs to θ^1, u_2 belongs to θ^2, and u_5 belongs to θ^2 and θ^3.

Assumption 1 allows to create $\Theta' = \left\{\theta'^1, \theta'^2, \theta'^3\right\}$, where $\theta'^1 = (u_1)$, $\theta'^2 = (u_2, u_5)$ and $\theta'^3 = (u_5)$. We do not necessarily recover θ in S but we recover a subset θ' instead.

Assumption 1 guarantees that we can identify all γ in S, this is because γ is no more than a combination of elements of θ, and if we can identify the elements of θ into the sample then we can identify any combination of them. Following with the previous example, let

$$\Gamma = \left\{\gamma^{11}, \gamma^{12}, \gamma^{13}, \gamma^{21}, \gamma^{22}, \gamma^{23}, \gamma^{24}, \gamma^{25}, \gamma^{26}, \gamma^{27}, \gamma^{31}, \gamma^{32}, \gamma^{33}\right\}$$

be the population of all possible subsets from each element of Θ, where $\gamma^{11} = (u_1)$, $\gamma^{12} = (u_3)$, $\gamma^{13} = (u_1, u_3)$, $\gamma^{21} = (u_2)$, $\gamma^{22} = (u_3)$, $\gamma^{23} = (u_5)$, $\gamma^{24} = (u_2, u_3)$, $\gamma^{25} = (u_2, u_5)$, $\gamma^{26} = (u_3, u_5)$, $\gamma^{27} = (u_2, u_3, u_5)$, $\gamma^{31} = (u_4)$, $\gamma^{32} = (u_5)$ and $\gamma^{33} = (u_4, u_5)$. Let $\Gamma^s = \left\{\gamma^{11}, \gamma^{21}, \gamma^{23}, \gamma^{25}, \gamma^{32}\right\}$ be the collection of all subsets γ in S.

The elements of Γ^s are identifiable into S because they are subsets of the elements of Θ', that is, $\Gamma^s = \cup_{\theta' \in \Theta'} \left\{P\left(\theta'\right) - \emptyset\right\}$. The identifiability of Γ^s by S is the reason why we create the artificial population Γ.

Assumptions 2 and 3 impose standard restrictions on the sampling designs.

Table 1 briefly summarizes the concepts and assumptions presented above.

[INSERT TABLE 1 HERE]

Briefly, we do not have a sample of the population whose parameters we are interested in, that is, we do not have a sample of Θ, but under Assumption 1 we have a sample of an "alternative" population Γ and we can estimate certain parameters of Γ which are linearly related with the parameters we are interested in.

We can summarize the proposed methodology in the following steps.

1. Transform the population of sets Θ into an artificial population of subsets, Γ. We know the total number of elements of a given size in both populations are linearly related.

2. Use S (a sample of U) to construct Γ^s (a sample of Γ). Assumption 1 allows to transform the sample S into the sample Γ^s.

3. Based on Γ^s estimate L estimating L_c for all c (see section below).

4. Once we estimate L, estimate T_k based on $T = A^{-1}L$.

3.2 Estimator for L_c

We can estimate L_c using the HT estimator as follows,

$$\hat{L}_c^{HT} = \sum_{\gamma \in \Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right], \tag{3}$$

where$I[\cdot]$ is the indicator function.

Lemma 1. *Under Assumptions 1 and 2, we have that* $E\left(\hat{L}_c^{HT}\right) = L_c$.

Proof. Let I_γ be an indicator function which takes the value one if $\gamma \in \Gamma^s$ and zero otherwise, then we have that $E\left(I_\gamma\right) = \pi_\gamma$. The estimator (3) can be expressed as $\hat{L}_c^{HT} = \sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] I_\gamma$, thus we have $E\left(\hat{L}_c^{HT}\right) = \sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] E\left(I_\gamma\right)$, and given that $E\left(I_\gamma\right) = \pi_\gamma$, we have that $\hat{L}_c^{HT}$ is an unbiased estimator of L_c. □

It is worth to note here that Assumption 2 only requires that any subset in Γ has positive probability to be sampled, it does not mean that the sample Γ^s must have elements γ of all different sizes. For $\hat{L}_c^{HT}$ be unbiased, it does not matter the size of the subsets γ in Γ^s, we only have to be sure that any subset could be sampled.

Let $p^s \leq p$ be the maximum size of γ in Γ^s, we have that $\hat{L}_c^{HT} = 0$ for all $c > p^s$, and also that $\hat{L}_c^{HT} > 0$ for all $c \leq p^s$. This statement is a consequence of the definition of Γ: if we have γ in the sample we also have all possible combinations of its elements, for instance, if we observe $\gamma^1 = (u_2, u_5)$ in Γ^s we also observe $\gamma^2 = (u_2)$ and $\gamma^3 = (u_5)$. In other words, if we observe γ we also observe every $\gamma' \subset \gamma$.

3.3 Estimator for L

We propose to estimate $L = (L_1, L_2, \ldots, L_p)'$ replacing each element by $\hat{L}_c^{HT}$, thus we have,

$$\hat{L}^{HT} = \left(\hat{L}_1^{HT}, \hat{L}_2^{HT}, \ldots, \hat{L}_p^{HT}\right)'. \tag{4}$$

Lemma 2. *Under Assumptions 1 and 2 we have that* $E\left(\hat{L}^{HT}\right) = L$.

Proof. Under Assumptions 1 and 2 $\hat{L}_c^{HT}$ is unbiased for all c, thus $\hat{L}^{HT}$ is unbiased. □

3.4 Variances and covariances for $\hat{L}_c^{HT}$

Let γ_c be an element of Γ of size c, and let $\gamma'_{c'}$ be an element of Γ of size c'. Then, $Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$ can be expressed as follows,

$$\sigma_{cc'} = Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_\gamma \pi_{\gamma'}\right).$$

It is worth to note that, for $c = c'$ the $Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = Var\left(\hat{L}_c^{HT}\right)$.

The proposed estimator for the covariance is given by

$$\hat{\sigma}_{cc'} = \widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = \sum_{\gamma\in\Gamma^s}\sum_{\gamma'\in\Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_\gamma \pi_{\gamma'}\right) \pi_{\gamma\gamma'}^{-1}.$$

Lemma 3. *Under Assumptions 1-3, we have that* $E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] = Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$.

Proof. See Appendix 8.2.1. □

Provided that the sample will not necessarily include subsets γ of all sizes, then $\hat{\sigma}_{cc'} = 0$ for $c \wedge c' > p^s$.

3.5 Variance-Covariance matrix for $\hat{L}^{HT}$

Let $\Omega_L = [\sigma_{cc'}]$ be the $(p \times p)$ variance-covariance matrix of $\hat{L}^{HT}$, replacing each element of Ω_L by its unbiased estimator we have the estimator for Ω_L, $\hat{\Omega}_L = [\hat{\sigma}_{cc'}]$, where $\hat{\sigma}_{cc'} = \widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$.

Lemma 4. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_L\right) = \Omega_L$

Proof. Under Assumptions 1-3, $\hat{\sigma}_{cc'}$ is unbiased for all c and c', thus $\hat{\Omega}_L$ is unbiased. □

4 The chained Horvitz-Thompson estimator for T

Based on equation (2) the estimator we propose for T is given by,

$$\hat{T}^{CHT} \quad = \quad A^{-1}\hat{L}^{HT}. \tag{5}$$

Basically, it consists in replacing L by $\hat{L}^{HT}$ in $T = A^{-1}L$.

Lemma 5. *Under Assumptions 1 and 2, we have that* $E\left(\hat{T}^{CHT}\right) = T$.

Proof. Under Assumptions 1 and 2 $\hat{L}^{HT}$ is unbiased , and given that A^{-1} is a non-stochastic matrix, $\hat{T}^{CHT}$ is also unbiased. □

The variance-covariance matrix of $\hat{T}^{CHT}$ is given by $\Omega_T = A^{-1}\Omega_L A'^{-1}$, our proposed estimator for Ω_T is given by

$$\hat{\Omega}_T \quad = \quad A^{-1}\hat{\Omega}_L A'^{-1}. \tag{6}$$

Lemma 6. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_T\right) = \Omega_T$.

Proof. Under Assumptions 1-3 $\hat{\Omega}_L$ is unbiased, and given that A^{-1} is a non-stochastic matrix, $\hat{\Omega}_T$ is also unbiased. □

4.1 Feasible estimator

We assume that we do not know p so the previous estimators are unfeasible because they depend on p (L and A^{-1} are of order p). As a special case, if we know p we use (5) and (6).

We assume that all we know about the size of the subsets γ is based on the sample, thus the largest size we observe is $p^s \leq p$. This fact could lead us to think that we can only estimate $L_{c\leq p^s}$, but this is not true because under Assumption 2 if we do not observe subsets of sizes $c > p^s$ is due to randomness, therefore $\hat{L}^{HT}_{c>p^s} = 0$ because $I\left[\#\gamma = c\right] = 0$ for $c > p^s$. The problem caused by not knowing p is that we do not know if there exist subsets of size $c > p^s$, and their estimates are zero.

Since the largest size of subsets we observe in the sample is p^s, then the estimate of the total number of sets of size larger than p^s is zero. Then $\hat{L}^{HT}_{c>p^s} = 0$ makes that $\hat{T}^{CHT}_{k>p^s} = 0$.

We write L, T and A as follows, $L = \left(L_A^{'}, L_B^{'}\right)^{'}$ where $L_A = \left(L_1, \ldots, L_{p^s}\right)^{'}$ and $L_B = \left(L_{p^s+1}, \ldots, L_p\right)^{'}$, $T = \left(T_A^{'}, T_B^{'}\right)^{'}$ where $T_A = \left(T_1, \ldots, T_{p^s}\right)^{'}$ and $T_B = \left(T_{p^s+1}, \ldots, T_p\right)^{'}$, and $A^{-1} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}$ where $\bar{A}_{11}$ is a $(p^s \times p^s)$ upper-left submatrix of A^{-1}, $\bar{A}_{12}$ is a $(p^s \times (p - p^s))$ upper-right submatrix of A^{-1}, $\bar{A}_{21} = [0]$ is a $((p - p^s) \times p^s)$ lower-left submatrix of A^{-1} and $\bar{A}_{22}$ is a $((p - p^s) \times (p - p^s))$ lower-right submatrix of A^{-1}.

Given that $T = A^{-1}L$ and the previous definitions, we have that $T_A = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} L$ and $T_B = \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} L$, given this we divide the estimator of T into two parts, on the one hand we estimate T_A and on the other hand we estimate T_B.

Using block matrix multiplication we have that,

$$\begin{aligned} \hat{T}^{CHT}_A &= \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} \hat{L}^{HT} \\ &= \begin{bmatrix} \bar{A}_{11}\hat{L}^{HT}_A & \bar{A}_{12}\ \hat{L}^{HT}_B \end{bmatrix} \\ &= \bar{A}_{11}\hat{L}^{HT}_A. \end{aligned} \tag{7}$$

The previous result is given by the fact that $\hat{L}_B^{HT} = 0$.

It is worth to note here that $\bar{A}_{11}$ is a submatrix of A^{-1}, if we need to know A^{-1} in order to get $\bar{A}_{11}$, the estimator $\hat{T}_A^{CHT}$ would be unfeasible because it would still depend on p. In appendix (10) we prove that $\bar{A}_{11} = A_{11}^{-1}$ where $A_{11} = A\left[1 : p^s, 1 : p^s\right]$, that is, $\bar{A}_{11}$ is the inverse of a submatrix of A given by its first p^s rows and columns, therefore, it is enough to know p^s to get $\bar{A}_{11}$.

In the same line

$$\begin{aligned} \hat{T}_B^{CHT} &= \left[\begin{array}{cc} \bar{A}_{21} & \bar{A}_{22} \end{array}\right] \hat{L}^{HT} \\ &= \left[\begin{array}{ccc} 0\hat{L}_A^{HT} & \bar{A}_{22} & \hat{L}_B^{HT} \end{array}\right] \\ &= 0. \end{aligned} \tag{8}$$

The previous result is given by $\bar{A}_{21} = 0$ and $\hat{L}_B^{HT} = 0$.

The key issue here is that, regardless of whether we know p, the estimate for T_B is equal to zero because its estimator depends entirely on $\hat{L}_B^{HT}$. In other words, the estimate of the total number of sets larger than p^s is zero, and the estimate of the total number of sets of sizes less than or equal to p^s is given by $\hat{T}_A^{CHT}$.

Summarizing, if we do not know p the only feasible estimator is $\hat{T}_A^{CHT}$. It does not mean we only have an estimator for T_A, we know that, if T_B exist, its (unbiased) estimate is zero.

Lemma 7. *Under Assumptions 1-2, we have that* $E\left(\hat{T}_A^{CHT}\right) = T_A$ *and* $E\left(\hat{T}_B^{CHT}\right) = T_B$.

Proof. Under Assumptions 1 and 2 $\hat{L}^{HT}$ is unbiased, and given that $\bar{A}_{11}$, $\bar{A}_{12}$, $\bar{A}_{21}$ and $\bar{A}_{22}$ are a non-stochastic matrices, $\hat{T}_A^{CHT}$ and $\hat{T}_B^{CHT}$ are also unbiased. □

Let Ω_{TA} be the variance-covariance matrix of $\hat{T}_A^{CHT}$, the estimator of Ω_{TA} is given by,

$$\hat{\Omega}_{TA} = A_{11}^{-1}\hat{\Omega}_{L,11}A_{11}^{-1'}, \tag{9}$$

where $\hat{\Omega}_{L,11} = \hat{\Omega}_L\left[1 : p^s, 1 : p^s\right]$ and $A_{11}^{-1} = \bar{A}_{11}$.

Let Ω_{TB} be the variance-covariance matrix of $\hat{T}_B^{CHT}$, the estimator of Ω_{TB} is given by $\hat{\Omega}_{TB} = [0]$ (we show this equality in Appendix 8.2.2).

Lemma 8. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_{TA}\right) = \Omega_{TA}$ *and* $E\left(\hat{\Omega}_{TB}\right) = \Omega_{TB}$.

Proof. See Appendix 8.2.2. □

4.2 Numerical Example

We present in this section a simple numerical example. Suppose we have a population of 10 houses grouped into 6 blocks in the following way: 3 blocks have 1 house ($T_1 = 3$), 2 blocks have 2 houses ($T_2 = 2$) and 1 block has 3 houses ($T_3 = 1$).

$U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$ is a population of $N = 10$ houses, $\Theta = \left\{\theta^1, \theta^2, \theta^3, \theta^4, \theta^5, \theta^6\right\}$ is a population of $J = 6$ blocks, where $\theta^1 = (u_1)$, $\theta^2 = (u_2)$, $\theta^3 = (u_3)$, $\theta^4 = (u_4, u_5)$, $\theta^5 = (u_6, u_7)$, $\theta^6 = (u_8, u_9, u_{10})$, the maximum size of sets (blocks) is $p = 3$.

The population of subsets is given by

$\Gamma = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, (u_4, u_5), (u_6, u_7), (u_8, u_9), (u_9, u_{10}), (u_8, u_{10}), (u_8, u_9, u_{10})\}$, the sizes of the elements of Γ are given by $c = 1, 2, 3$. Each element in Γ is a possible combination of elements of each set in Θ.

We draw a random sample S of size $n = 4$ from U. The probability of selection of γ_c is given by

$$Prob\left(\gamma_c \in S\right) = \left(\begin{array}{c} N \\ n \end{array}\right)^{-1} \left(\begin{array}{c} N-c \\ n-c \end{array}\right).$$

Given that $c = 1, 2, 3$ we have that $Prob\left(\gamma_1 \in S\right) = \frac{n}{N}$, $Prob\left(\gamma_2 \in S\right) = \frac{n(n-1)}{N(N-1)}$ and $Prob\left(\gamma_3 \in S\right) = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}$.

Let $S = \{u_1, u_3, u_6, u_7\}$ be a sample of U, and let $\Gamma^s = \{u_1, u_3, u_6, u_7, (u_6, u_7)\}$be the sample of Γ that we get from S. The size of largest subset in Γ^s is $p^s = 2 < p$.

We first estimate $L = (L_A,\, L_B)$ by (4), where $L_A = (L_1,\, L_2)$ and $L_B = (L_3)$, we have that

$$\hat{L}_1^{HT} = 4\pi_{\gamma 1}^{-1} = 4\left(\frac{4}{10}\right)^{-1} = 10$$

and

$$\hat{L}_2^{HT} = 1\pi_{\gamma 2}^{-1} = 1\left(\frac{12}{90}\right)^{-1} = 7.5$$

with the previous we have that $\hat{L}_A^{HT} = (10,\, 7.5)'$.

We are assuming we do not know p, so we neither know that L_B exist, that is, we do not know there are subsets of size $c = 3$ in Γ but, under assumption 2, we know that if there are subsets of size larger than $p^s = 2$ the estimate of their total number is zero, that is, $\hat{L}_3^{HT} = 0$. Assumption 2 establishes that any subset in Γ should have positive probability to be sampled, given the size of the largest subset in Γ is 3, as long as $n \geq 3$ Assumption 2 holds.

On the other hand we have $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$ so $A_{11}^{-1} = (A\,[1:2,\, 1:2])^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}$. Given the previous we have that

$$\hat{T}_A^{CHT} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \hat{L}_A^{HT},$$

thus we have that $\hat{T}_A^{CHT} = (-5,\, 7.5)'$.

As we see, in theory some T_k could be negative, but in practice they should be positive. We can impute the value zero for every $\hat{T}_k^{CHT} < 0$, this generate some bias but reduce the Mean Square Error.

We do not have an estimate of T_3 because we do not know there are subsets of size 3, but we know that the estimate of every $T_{k>2}$ is zero (if it exist!).

5 Estimating the network degree distribution

In this section we use the methodology proposed above to estimate the network degree distribution, which is one of the most fundamental features of a network. Some of the papers related with this topic are the following.

5.1 Literature review

Frank (1980, 1981) shows that, under certain networks sampling designs, the expectation of the observed degree relative frequencies (i.e., the degree distribution) is a linear combination of the true degree relative frequencies, and he proposes an estimator which depend on an inverse matrix that in some cases is not invertible and, even when it is, the result may not be non-negative.

Zhang et al. (2015) proposes a method to overcome these problems, and apply it to a few common networks sampling designs where inclusion probabilities are known. They do not make assumptions about the structure of the network. The methodology is assessed by a simulation study that considers the effects of several factors on the accuracy of the estimators. The results show, among others aspects, that sampling schemes have a considerable impact on the performance of the estimators.

Thompson (2006) proposes a sampling design and discusses inference procedures on the average degree and the degree distribution under such sampling designs. In the same line, Ribeiro & Towsley (2012) studies the mean squared error associated with different sampling methods for the degree distribution.

Stumpf & Wiuf (2005) discusses two sampling schemes for selecting random subnets from a network and investigate how the degree distribution is affected for this two types of sampling. The central question addressed by the authors is whether the degree distribution of randomly sampled subnets has the same properties as the degree distribution of the overall network, they derive a necessary and sufficient condition that guarantees this equality and describe some situations under which this condition is satisfied, however, for the majority of the networks this condition is no be met.

Internet topology is one of the areas where networks are widely used. Faloutsos et al. (1999) shows that the internet degree distribution has a power-law form, Lakhina et al. (2003) show that when graphs are sampled using traceroute-like methods, the resulting degree distribution (based on sampled network) can differs sharply from the true, they explore the reason of such bias and propose a test for determining when sampling bias is present. Achlioptas et al. (2009) study the traceroute sampling systematically and extend the results in Lakhina et al. (2003).

5.2 Some network definitions

A common way to represent a network is listing all their nodes and links among them. Let $\mathcal{N} = \{1, 2, ..., N\}$ be a collection of N nodes and let g be a collection of links. Then a network is defined by $\mathcal{G} = (\mathcal{N}, g)$. If $ij \in g$, then node i is linked to node j, alternatively, we can use the notation $g_{ij} = 1$ if nodes i and j are linked, $g_{ij} = 0$ otherwise.

The relation among nodes can be directed or undirected. In directed networks if i is linked to j, j is not necessarily linked to i, that is, $g_{ij} \neq g_{ji}$ is possible. In undirected networks there exists reciprocity, and then, if i is linked to j, j is linked to i as well, $g_{ij} = g_{ji}$.

The degree of a node i, $d_i(\mathcal{G})$, is the number of links i has. The degree distribution of a network, $P_{\mathcal{G}}$, is a description of the relative frequencies of nodes that have different degrees. Let $T_k = \sum_{i=1}^{N} I[d_i(\mathcal{G}) = k]$ the total number of nodes with degree k in $\mathcal{G}$, we have that T_k/N is the relative frequency of nodes with degree k in $\mathcal{G}$. Knowing T_k for all k, we can get the degree distribution of $\mathcal{G}$.

Let $S \subset N$ be a subset of nodes, we name $\mathcal{G}_S = (S, g|_S)$ to the network $\mathcal{G}$ restricted only to the set S. That is,

$$g|_{Sij} = \begin{cases} 1 & \text{if } i\,j \in S,\, g_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}.$$

In the network literature there exist multiple networks sampling designs. We only work with two of them known as *induced subgraph* and *incident subgraph*. In induced subgraph sampling a set of nodes is selected and then all edges among selected nodes are observed. In incident subgraph links are selected and then all nodes that correspond to selected links are observed. The selection can be done under simple random sampling or under Bernoulli sampling.

5.3 Estimating network degree distribution by CHT estimator

In this subsection we apply the CHT estimator for estimating the network degree distribution. The proposed estimator can be used under fairly general contexts, the main requirement being to have a probability sample of links.

As noted above the CHT estimator works in a particular framework. This framework is given

by the existence of a population U whose elements are grouped into sets in Θ, the main goal of the CHT estimator is to estimate the total number of sets of a given size in Θ based on a sample of U. In order to apply the CHT estimator we need to adapt some networks concepts to the framework established in Sections 2 and 3.

5.3.1 Definitions

We begin by defining a network, a population of links and a sample of links.

Definition 7. Let $\mathcal{G} = (\mathcal{N}, g)$ be a fixed finite network with $\mathcal{N} = \{1, 2, ..., N\}$ nodes. Let $U = g$ be a population given by the links of $\mathcal{G}$, and let $g|_S$ be a probability sample of U .

Definition 8. Let $\theta^i = \{g_{ij} | \, g_{ij} = 1, \, for \, j \in \mathcal{N}\}$ be the set of all links that i has, with $i = 1, \ldots, J$ and $j = 1, \ldots, N$, where J is the total number of nodes in $\mathcal{G}$ with at least one link.

The *neighborhood* of a node i is the set of all nodes that i is linked to, $N_i(g) = \{j : g_{ij} = 1\}$. The set θ^i is conceptually close to $N_i(g)$, the latter contains ***the nodes*** linked to i while the former contains ***the links*** i has. It is important to note here that, for $d_i(\mathcal{G}) \neq 0$, $d_i(\mathcal{G}) = \#\theta^i$, the degree of the node i is equal to the total number of elements in θ^i.

Definition 9. Let $\Theta = \cup_{i=1}^{J} \theta^i$, be the collection of $\theta' s$.[1]

Given the previous definitions, the total number of sets of size k in Θ is equal to the total number of nodes with degree k in $\mathcal{G}$, therefore $T_k = \sum_{i=1}^{J} I\left[\#\theta^i = k\right]$. Our main goal here is to estimate T_k.

Up to here we defined a population U whose elements (links) are grouped into sets which belong to Θ. At this point it is worth to remember that the methodology proposed in Section 3 allows to estimate the total number of sets of a given size based on a sample of elements from the sets instead of a sample of sets, so based on a sample of U (links) we can estimate the total number of sets of size k in Θ, what is no more than the total number of nodes with degree k in $\mathcal{G}$. In other words, based on a sample of links we can estimate the total number of nodes with degree k in $\mathcal{G}$.

Next we define an "artificial" population constructed from Θ.

Definition 10. Let $\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$ be a collection of all possible subsets we can get from ***each element*** of Θ, where $P(\theta)$ is the power set of θ, γ is a generic element of Γ and γ_c a generic

[1] We define Θ as a collection because their elements are not necessarily exclusive.

element of size c, with $c = \{1, 2, \ldots, p\}$. The total number of elements of size c in Γ is given by $L_c = \sum_{\gamma \in \Gamma} I\left[\#\gamma = c\right]$.

The elements of θ are links then, given a node i, the collection Γ contains all combinations of one link of i, all combinations of two links of i, all combinations of three links, etc.[2], and this is so for all i.

Definition 11. Let $\Gamma^s = \{\gamma | \forall \gamma \in g|_S\}$ be the collection of all subsets γ in $g|_S$. Let $p^s \leq p$ be the maximum size of $\gamma \in g|_S$.

$\gamma \subseteq \theta$ are subsets of links, what we mean by $\gamma \in g|_S$ is that all links in γ are in $g|_S$. By now we do not say anything about $g|_S$, it is only a probability sample of links.

Definition 12. Let $\pi_\gamma = Prob\left(\gamma \in \Gamma^s\right)$ be the probability of selection of γ and $\pi_{\gamma\gamma'} = Prob\left(\gamma \cup \gamma' \in \Gamma^s\right)$ be the joint probability of selection of γ and γ'. Given definition 11, $\pi_\gamma = Prob\left(\gamma \in g|_S\right)$ and $\pi_{\gamma\gamma'} = Prob\left(\gamma \cup \gamma' \in g|_S\right)$.

Given that γ is a subset of links, the probability of selection of γ is the joint probability of selection of such links.

5.3.2 Assumptions

We are going to use the CHT estimator for estimating the network degree distribution. The unbiasedness and feasibility of the estimator require assumptions 1 and 2 hold, and the unbiasedness of the variance estimator requires assumptions 1-3 hold. We present below these assumptions adapted to the network context.

First we assume some mild conditions under which Assumption 1 holds.

Assumption 4. *Let $g|_S$ be a probability sample of links from $\mathcal{G}$. For all $[g|_S]_{ij} = 1$ we observe i.*

Let $[g|_S]_{ij}$ be a sampled link between i and j. Assumption (4) establishes we know such link belongs to node i. This requirement makes feasible to group the sampled links as they are grouped in θ so we have a sample of subsets of θ. The "artificial" population Γ contains all possible subsets of every θ, therefore we have a sample of Γ, that is Γ^s.

The next two assumption refer to the individual and joint probability of selection of γ.

[2] Assuming that $d_i\left(G\right) \geq 3$.

Assumption 5. $\pi_\gamma > 0 \; \forall \gamma \in \Gamma$. *The probability of selection of any element in* Γ *has to be positive.*

Assumption 6. $\pi_{\gamma\gamma'} > 0 \; \forall \gamma \vee \gamma' \in \Gamma$. *The joint probability of selection of any pair of elements in* Γ *has to be positive.*

Assumptions 5 and 6 are equivalent to assumptions 2 and 3. The elements in Γ are subsets of links that belong to the same node, therefore we need that any set of links that belongs to the same node has positive probability to be sampled (individual probability), and not only that but also we need that any combination of two sets of links has positive probability to be sampled (joint probability).

5.3.3 Estimator

Let $\tilde{T} = \left(T_0, T'\right)'$ be a $((p+1) \times 1)$ vector with $T = (T_1, T_2, \ldots, T_p)'$, where T_0 and T_k are the total number of nodes with degree zero and k in $\mathcal{G}$ respectively[3], therefore we have that $P_\mathcal{G} = N^{-1}\tilde{T}$. Following we present an estimator for $P_\mathcal{G}$ based on the methodology proposed in the previous section.

Based on definitions (7), (8) and (9) we have a population of links named U whose elements are grouped into sets which belongs to Θ. We know that the total number of sets of size $0 < k \leq p$ in Θ is equal to the total number of links with degree k in $\mathcal{G}$, that is, T_k. We propose to use the methodology presented above for estimating T_k.

An important point to note here is we assume p is unknown, that is, we do not know the maximum degree in $\mathcal{G}$, therefore, following the exposed in section (4.1), we only going to have estimates for $T_{k \leq p^s}$, but it does not mean we can only estimate T partially, because we know that the estimates for $T_{k > p^s}$ are zero.

Being L_c the total number of subsets of size c in Γ, the HT estimator for L_c is given by

$$\hat{L}_c^{HT} = \sum_{\gamma \in \Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right]. \tag{10}$$

Given we are assuming p is unknown we propose to use the estimator (7) for estimating T as follows

[3]We divide the vector $\tilde{T}$ into T_0 and T because our methodology allows to estimate T, that is, the total number of nodes with degree $k > 0$. For estimating the total number of nodes with degree zero, T_0, we present another estimator.

$$\hat{T}_A^{CHT} = \bar{A}_{11}\hat{L}_A^{HT}, \tag{11}$$

where $\hat{L}_A^{HT} = \left(\hat{L}_1^{HT}, \ldots, \hat{L}_{p^s}^{HT}\right)'$ and $\hat{T}_A^{CHT} = \left(\hat{T}_1^{CHT}, \ldots, \hat{T}_{p^s}^{CHT}\right)'$.

Despite the previous estimator only estimates $T_{k \leq p^s}$, we know that, if there exist nodes with degree larger than p^s the estimate of their total number is zero.

Up to here we estimate $T_{k \geq 1}$, we do not have an estimator for T_0, that is, we do not have an estimator for the total number of nodes without connections, but given that $T_0 = N - \sum_{k=1}^{p} T_k$, assuming we know N we can estimate T_0 by $\hat{T}_0^{CHT} = N - 1_{p^s}\hat{T}_A^{CHT}$ where 1_{p^s} is a $(1 \times p^s)$ vector of ones. Here we are assuming we know N, but if we do not, N should be replaced by any unbiased estimator.

Our proposed estimator for the network degree distribution is given by

$$\hat{P}_{\mathcal{G}}^{CHT} = N^{-1}\hat{\tilde{T}}_A^{CHT}, \tag{12}$$

where $\hat{\tilde{T}}_A^{CHT} = \left(\hat{T}_0^{CHT}, \hat{T}_A^{CHT'}\right)'$.

Given we are assuming p is unknown, we do not know if there are nodes with degree greater than p^s, but we know if they exist the estimate of their total number is zero, so the estimate of the relative frequencies of nodes with degree greater than p^s is zero.

The $Var\left(\hat{T}_0^{CHT}\right) = Var\left(N\right) + Var\left(1_{p^s}\hat{T}_A^{CHT}\right) - 2Cov\left(N, 1_{p^s}\hat{T}_A^{CHT}\right)$, and given that N is not random, we have that $Var\left(\hat{T}_0\right) = Var\left(1_{p^s}\hat{T}_A^{CHT}\right)$. The estimator of $Var\left(\hat{T}_0^{CHT}\right)$ is given by $\widehat{Var}\left(\hat{T}_0^{CHT}\right) = 1_{p^s}\hat{\Omega}_{TA}1_{p^s}'$.

The variance estimator of $\hat{P}_{\mathcal{G}}^{CHT}$ is given by

$$\widehat{Var}\left(\hat{P}_{\mathcal{G}}^{CHT}\right) = N^{-2}\left(\widehat{Var}\left(\hat{T}_0^{CHT}\right), Diag\left(\hat{\Omega}_{TA}\right)\right).$$

5.4 Estimating degree distribution under induced and incident subgraph sampling designs. Probabilities

All of the previous developments are valid for any probability sampling design. Here, we derive the probabilities of selection under induced and incident subgraph sampling for directed networks. The extension to undirected networks is straightforward.

5.4.1 Probability of selection under induced subgraph sampling

Under induced subgraph sampling we draw a sample of nodes and we know their relation (we know the links among sampled nodes), therefore we have to construct the probability of selection of links based on the probability of selection of nodes.

The probabilities of selection of a link is given by the joint probability of selection of the two nodes involved in such link. It is important to remember that γ is a subset of links that belong to the same node, therefore all links in γ have a common node, thus the total number of nodes involved in γ is equal to the total number of links in γ plus one. Thus, the probability of selection of γ_c is given by the joint probability of selection of the $(c+1)$ nodes involved in γ_c.

$$Prob\left(\gamma_c \in g|_S\right) = \left(\begin{array}{c} N \\ n \end{array} \right)^{-1} \left(\begin{array}{c} N-(c+1) \\ n-(c+1) \end{array} \right),$$

where n is the number of sampled nodes, $\left(\begin{array}{c} N \\ n \end{array} \right)$ are all possible samples of size n, and $\left(\begin{array}{c} N-(c+1) \\ n-(c+1) \end{array} \right)$ are all possible samples of size n which contain all the nodes involved in the c links of γ.

On the other hand, $\pi_{\gamma\gamma'}$ is the joint probability of selection of γ and γ', this is equal to the joint probability of selection of all nodes involved in the links of γ and γ'. The nodes involved in the links of γ and $\gamma^{'}$ could be repeated, for instance, being g_{ij} a link in γ, the link g_{ji} could be in γ', thus both links are formed by the same two nodes.

Let $\omega_{\gamma\gamma'} = \{i, j | \, g_{ij} \in \gamma \cup \gamma'\}$ be a set of all nodes (without repetitions)[4] involved in the links of γ and γ', we have

[4]Since $\omega_{\gamma\gamma'}$ is an union of nodes, if there are some common nodes among the links in γ and γ' they will appear only once in $\omega_{\gamma\gamma'}$.

$$Prob\left(\left(\gamma_c \cup \gamma_{c'}^{'}\right) \in g|_S\right) \quad = \quad \left(\begin{array}{c} N \\ n \end{array}\right)^{-1} \left(\begin{array}{c} N - \omega_{\gamma\gamma'} \\ n - \omega_{\gamma\gamma'} \end{array}\right).$$

In case the links involved in γ_c and $\gamma_{c'}^{'}$ do not have nodes in common, $\#\omega_{\gamma\gamma'} = (2c+2)$, and in case the links involved in γ and γ' have all nodes in common, $\#\omega_{\gamma\gamma'} = (c+1)$.

In order to the estimator (12) be unbiased it is necessary that $\pi_\gamma > 0$ for all γ and this happens as long as $n \geq (c+1)$ for all c, and given $c = \{1, 2, \ldots, p\}$ this implies that the sample size should be, at least, equal to the number of nodes involved in the links of γ_p, that is, $p+1$.

Similarly, in order to the estimator of the variance be unbiased, in addition to the previous assumption on π_γ, it is necessary to assume that $\pi_{\gamma\gamma'} > 0$ for allγ and γ', and this happens as long as $n \geq \#\omega_{\gamma\gamma'}$ for all γ and γ'. This implies that the sample size should be, at least, equal to the number of nodes involved in the links of all pairs of subsets γ and γ', therefore, in order to $\pi_{\gamma\gamma'} > 0$ we need that $n \geq max\left(\#\omega_{\gamma\gamma'}\right)$.

5.4.2 Probability of selection under incident subgraph sampling

Under incident subgraph sampling we sampled links directly, the probability of selection of γ is given by the joint probability of selection of the links involved in γ, and this is given by

$$Prob\left(\gamma_c \in g|_S\right) \quad = \quad \left(\begin{array}{c} M \\ m \end{array}\right)^{-1} \left(\begin{array}{c} M - c \\ m - c \end{array}\right),$$

where M is the total number of links in the network and m is the sample size (the total number of sampled links).

Since each link corresponds to a unique node, the joint probability of selection of γ and γ' is given by the joint probability of selection of the $(c + c')$ links involved in γ_c and $\gamma_{c'}^{'}$.

$$Prob\left(\left(\gamma_c \cup \gamma_{c'}^{'}\right) \in g|_S\right) \quad = \quad \left(\begin{array}{c} M \\ m \end{array} \right)^{-1} \left(\begin{array}{c} M-(c+c') \\ m-(c+c') \end{array} \right).$$

Under this sampling design, π_γ will be greater than zero for all γ_c as long as $m \geq c$ for all c, therefore $m \geq p$, and $\pi_{\gamma\gamma'}$ will be greater than zero for all pairs of sets γ_c and $\gamma_{c'}^{'}$ as long as $m \geq (c+c')$ for all c.

Summarizing, regardless of the sampling designs, in order to make the CHT estimator unbiased, *all links* involved in any node should have positive probability of being jointly sampled, and in order to make the variance estimator unbiased, *all links involved in any pairs of nodes* should have positive probability of being jointly sampled.

6 Monte Carlo simulations

In order to evaluate the performance of the CHT estimator for degree distribution we present Monte Carlo simulations.

We simulate the coordinates of N points (nodes) by two $U(0,1)$ independent random variables, and then we establish links among points following the p-nearest neighbors criteria, and construct the corresponding adjacency matrix. The networks created in this way ensure that each node has degree p (where p is the number of neighbors), i.e., the so called "regular networks" (networks whose nodes have all the same degree).

Second, in order to construct networks whose nodes have different degrees we randomly delete some links by taking rows and columns from the adjacency matrix and filling them with zeros. We randomly select the id's of the rows and columns to fill with zeros drawing two random samples from $\{1,\ldots,N\}$. We draw one sample of size $\lceil \alpha_1 N \rceil$ for selecting rows and other of size $\lceil \alpha_2 N \rceil$ for selecting columns, with $0 < \{\alpha_1, \alpha_2\} < 1$. In this way we have networks whose nodes have different degrees but none of them are greater than p.

We consider networks of two different sizes, $N = 300, 600$, whereas the sampling rates are $\{0.9, 0.7, 0.5\}$. The maximum degree is fixed in $p = 4$, so the degree distribution has domain in $k = \{0, 1, \ldots, 4\}$.

We want to ensure that the relative frequencies of nodes with degree k are not smaller than 0.1, that is, we want to ensure that $P_{\mathcal{G}}(d=k) > 0.1$ for all k. For doing that we fix $\alpha_1 = 0.1$ and $\alpha_2 = 0.4$.

Given a sample size, we generate one network and draw $R = 500$ samples for each sampling rate. The networks are the same at each repetition for all sampling rates and for both sampling schemes.

For induced subgraph sampling we randomly choose n numbers from $\{1, \ldots, N\}$ and select the nodes that match with such numbers. Under this sampling design we know all links among sampled nodes, so we can construct an adjacency matrix for the sampled nodes. Once we have the links and the probability of selection (previously presented) we construct the CHT estimator and its variance estimator.

For incident subgraph sampling we list the M links of the network and then we randomly choose m numbers from $\{1, \ldots, M\}$ and select the links whose row number match with such numbers. Under this sampling design we sampled links but we know the nodes involved in such links, so we can construct an adjacency matrix for the (indirectly) sampled nodes and with the probability of selection we construct the CHT estimator and its variance estimator. For this case, we know both M and N (i.e., the number of nodes) in order to estimate the number of nodes with degree 0. Therefore, the average degree is known.

In both sampling designs we have a sample of nodes for which we know the relation among them (links). We group all the links (within the sample) incident to a given node and then we construct all the different subsets we can make based on it. We do this for each sampled node, these subsets of links are the subsets γ in $g|_S$.

We assume that we do not know p, so the estimator for $T_{k>p^s}$ is zero. We take into account such zeros when we construct the estimators.

Depending on the sampling design we report the sampling rates (n/N) or (m/M). For each Monte Carlo exercise we report the true degree distribution, $P_{\mathcal{G}}(d=k)$ and the true average degree. , $ave.P_{\mathcal{G}}$. We report the estimates for the average degree only under induced subgraph sampling because under incident subgraph sampling the average degree is known, i.e., M/N. Then we also

report the bias and root-mean squared error (RMSE) in each case, that is,

$$Bias\left(d=k\right)=\frac{1}{R}\sum_{r=1}^{R}\left(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-P_{\mathcal{G}}\left(d=k\right)\right),$$

$$Bias\left(ave\right)=\frac{1}{R}\sum_{r=1}^{R}\left(ave.\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-ave.P_{\mathcal{G}}\left(d=k\right)\right),$$

$$RMSE\left(d=k\right)=\left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-P_{\mathcal{G}}\left(d=k\right)\right)^{2}\right]^{1/2},$$

$$RMSE\left(ave\right)=\left[\frac{1}{R}\sum_{r=1}^{R}\left(ave.\hat{P}_{\mathcal{G}}^{r}-ave.P_{\mathcal{G}}\right)^{2}\right]^{1/2},$$

respectively. We also report the average of the estimated standard deviations, $sd\left(d=k\right)=\frac{1}{R}\sum_{r=1}^{R}\hat{sd}(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right))$ and $sd(ave)=\frac{1}{R}\sum_{r=1}^{R}\hat{sd}(ave.\hat{P}_{\mathcal{G}}^{r})$ in order to assess the proposed estimator of the variance.

We compare our proposed estimator with the "naïve" one, where the in-sample degree is calculated. We also report the bias and RMSE for this case.

Tables 2 and 3 report the simulation results for the induced sampling for $N=300$ and $N=600$ network sizes, respectively. Tables 4 and 5 show the results for the incident sampling for $N=300$ and $N=600$ network sizes, respectively.

[INSERT TABLE 2 HERE]

[INSERT TABLE 3 HERE]

[INSERT TABLE 4 HERE]

[INSERT TABLE 5 HERE]

The simulation results highlight that the naïve estimator of the degree distribution that uses the sampled network as if it were the true one produces considerable bias in the estimation of both, the degree distribution and the average degree. The bias increases when the sampling rate decreases, and the same problem arises for both types of sampling designs. Our proposed CHT estimator, has a small bias in all cases. The CHT estimator also has a good performance in terms of RMSE.

The average of the estimated standard deviation is close to the simulated RMSE, and this suggests the estimator of the variance performs well in small samples.

With respect to the sampling rates and the population sizes the results are expected. The greater

the sampling rate is, the better the estimators behave, and in the same line, given a sampling rate, the greater the population is, the better the estimators behave. We cannot compare the results between the two sampling schemes because the sampling rates are not comparable, e.g., sampling 50% of the links is different from sampling 50% of the nodes.

7 Conclusion

This paper presents a general methodology for estimating the total number of sets of a given size based on a sample of elements of such sets. The resulting estimator is the chained Horvitz-Thompson (CHT) estimator.

We apply such methodology to derive an estimator for the degree distribution (and its variance) to be used under probability sampling designs, and we adapt it to two widely used sampling designs known as induced subgraph and incident subgraph. The results obtained from the Monte Carlo simulations show that the estimators perform well in terms of bias and RMSE, in line with the analytical results.

This is a basal paper, and there are many interesting lines in which to continue working. It would be worth to derive the asymptotic distribution of the CHT estimator. This should be relatively straightforward because the CHT estimator is no more than a linear combination of the HT estimator, and under some regularity conditions such estimator is asymptotically normal.

It would also be interesting to adapt the degree distribution estimator to other probability sampling designs and to evaluate them under other schemes. In the same line, more structure could be added to the degree distribution estimator. By assuming some particular degree distribution (scale-free, Poisson, etc.) the results may improve.

8 Appendix

8.1 Some properties of A and A^{-1}

The $((p+1)\times(p+1))$ Pascal Matrix $P=[p_{ij}]$ is defined by $p_{ij}=\begin{pmatrix} j \\ i \end{pmatrix}$ with $i,\, j=0,1,\ldots,p$.

The inverse $P^{-1}=\left[\bar{p}_{ij}^{-1}\right]$ is defined by $\bar{p}_{ij}^{-1}=(-1)^{j+i}\begin{pmatrix} j \\ i \end{pmatrix}$.

Lemma 9. *Let $A=P\left[1:p,\,1:p\right]$ be a $(p\times p)$ submatrix of P, then $A^{-1}=P^{-1}\left[1:p,\,1:p\right]$.*

Let $P_n=\begin{bmatrix} B & C \\ D & A \end{bmatrix}$, where $B_{(1\times1)}=[1]$, $C_{(1\times p)}=[1,\ldots,1]$ and $D_{(p\times1)}=[0,\ldots,0]^{'}$. By block matrix inversion we have that

$$\begin{bmatrix} B & C \\ D & A \end{bmatrix}^{-1}=\begin{bmatrix} B^{-1}+B^{-1}C\left(A-DB^{-1}C\right)^{-1}DB^{-1} & -B^{-1}C\left(A-DB^{-1}C\right)^{-1} \\ -\left(A-DB^{-1}C\right)^{-1}DB^{-1} & \left(A-DB^{-1}C\right)^{-1} \end{bmatrix}$$

Given $D_{(p\times1)}=[0,\ldots,0]^{'}$ we have

$$\begin{bmatrix} B & C \\ D & A \end{bmatrix}^{-1}=\begin{bmatrix} B^{-1} & -CA^{-1} \\ D & A^{-1} \end{bmatrix}$$

The inverse of B and A exists.

Given the previous we have that $P^{-1}\left[1:p,\,1:p\right]=A^{-1}$, and thus $\bar{a}_{ij}^{-1}=(-1)^{j+i}\begin{pmatrix} j \\ i \end{pmatrix}$ with $i,\, j=1,\ldots,p$.

The matrix P has a well known inverse, and given that A is a submatrix of P, we proved above that the inverse of$A$ is a submatrix of the inverse of P.

Lemma 10. *Let $A_{11}=A\left[1:p^s,\,1:p^s\right]$ be a $(p^s\times p^s)$ upper-left submatrix of A with $p^s\leq p$, then $A_{11}^{-1}=A^{-1}\left[1:p^s,\,1:p^s\right]$.*

$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ where A_{11} is a $(p^s\times p^s)$ upper-left submatrix of A, A_{12} is a $(p^s\times(p-p^s))$ upper-right submatrix of A, $A_{21}=[0]$ is a $((p-p^s)\times p^s)$ lower-left submatrix of A and A_{22} is a $((p-p^s)\times(p-p^s))$ lower-right submatrix of A.

By block matrix inversion we have that

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \\ -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix},$$

given $A_{21} = [0]$ we have

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ A_{21} & A_{22}^{-1} \end{bmatrix},$$

A_{11} and A_{22} are unitriangular matrices so their inverses exist.

Given the previous we have that $A^{-1}[1:p^s, 1:p^s] = A_{11}^{-1}$, that is, the inverse ofA_{11} is the $(p^s \times p^s)$ upper-left submatrix of A^{-1}. It is so relevant because this implies we do not need to know A to get A_{11}^{-1}

8.2 Variance and covariance

As we saw, the estimator $\hat{L}_c^{HT}$ can be expressed as $\hat{L}_c^{HT} = \sum_{\gamma\in\Gamma} \pi_\gamma^{-1} I[\#\gamma = c] I_\gamma$, the first indicator select the elements γ of size c in Γ and the second indicator identifies which are in the sample. In the same way we have $\hat{L}_{c'}^{HT} = \sum_{\gamma'\in\Gamma} \pi_{\gamma'}^{-1} I[\#\gamma' = c'] I_{\gamma'}$.

$$\begin{aligned} Cov\left(\hat{L}_c^{HT}, \hat{L}_{c'}^{HT}\right) &= Cov\left(\sum_{\gamma\in\Gamma} \pi_\gamma^{-1} I[\#\gamma = c] I_\gamma, \sum_{\gamma'\in\Gamma} \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] I_{\gamma'}\right) \\ &= \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I[\#\gamma = c] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] Cov\left(I_\gamma, I_{\gamma'}\right) \\ &= \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I[\#\gamma = c] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left[E\left(I_\gamma I_{\gamma'}\right) - E\left(I_\gamma\right) E\left(I_{\gamma'}\right)\right]. \end{aligned}$$

The $E(I_\gamma) = \pi_\gamma$, the expectation of γ be sampled is equal to the probability of γ be sampled, and the same is for $E\left(I_{\gamma'}\right) = \pi_{\gamma'}$. On the other hand $E\left(I_\gamma I_{\gamma'}\right) = \pi_{\gamma\gamma'}$, where $\pi_{\gamma\gamma'}$ is the joint probability of selection of γ and γ'.

Then using the previous results we have

$$Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_{\gamma}^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_{\gamma}\pi_{\gamma'}\right).$$

8.2.1 Variance-covariance estimator for L_c

The covariance estimator proposed in (3.4) can be written as

$$\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_{\gamma}^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_{\gamma}\pi_{\gamma'}\right) \pi_{\gamma\gamma'}^{-1} \left(I_{\gamma} I_{\gamma'}\right),$$

where $\left(I_{\gamma} I_{\gamma'}\right)$ is an indicator function that takes the value one if $\gamma \wedge \gamma' \in \Gamma^s$ and zero otherwise. With this we have

$$E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_{\gamma}^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_{\gamma}\pi_{\gamma'}\right) \pi_{\gamma\gamma'}^{-1} E\left(I_{\gamma} I_{\gamma'}\right),$$

with $E\left(I_{\gamma} I_{\gamma'}\right) = \pi_{\gamma\gamma'}$, and thus $E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] = Cov\left(\hat{T}_c^{HT},\, \hat{T}_{c'}^{HT}\right)$.

8.2.2 Variance-covariance matrix estimator for $\hat{T}_A^{CHT}$ and $\hat{T}_B^{CHT}$

By equation (7), the variance-covariance matrix of $\hat{T}_A^{CHT}$ is given by $\Omega_{TA} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} \Omega_L \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix}'$. Replacing Ω_L by $\hat{\Omega}_L$ we have the estimator for Ω_{TA} . Under assumptions 1-3 $\hat{\Omega}_L$ is unbiased, therefore, $\hat{\Omega}_{TA}$ it is also.

By equation (8), the variance-covariance matrix of $\hat{T}_B^{CHT}$ is given by $\Omega_{TB} = \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \Omega_L \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}'$. Replacing Ω_L by $\hat{\Omega}_L$ we have the estimator for Ω_{TB} . Under assumptions 1-3 $\hat{\Omega}_L$ is unbiased, therefore, $\hat{\Omega}_{TB}$ it is also.

To show that the previous expressions can be reduced we divide $\hat{\Omega}_L$ into four submatrix as follows, $\hat{\Omega}_{L,11} = \hat{\Omega}_L\left[1 : p^s,\, 1 : p^s\right]$, $\hat{\Omega}_{L,12} = \hat{\Omega}_L\left[1 : p^s,\, (p^s+1) : p\right]$, $\hat{\Omega}_{L,21} = \hat{\Omega}_L\left[(p^s+1) : p,\, 1 : p^s\right]$ and $\hat{\Omega}_{L,22} = \hat{\Omega}_L\left[(p^s+1) : p,\, (p^s+1) : p\right]$.

Given that $\hat{\sigma}_{cc'} = 0$ for $c \wedge c' > p^s$ we have that $\hat{\Omega}_{L,12} = [0]$, $\hat{\Omega}_{L,21} = [0]$ and $\hat{\Omega}_{L,22} = [0]$, therefore

$$\begin{aligned}\hat{\Omega}_{TA} &= \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} \begin{bmatrix} \hat{\Omega}_{L,11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix}' \\ &= A_{11}^{-1} \hat{\Omega}_{L,11} A_{11}^{-1'}.\end{aligned}$$

given that $\bar{A}_{11} = A_{11}^{-1}$.

$$\begin{aligned}\hat{\Omega}_{TB} &= \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} \hat{\Omega}_{L,11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}' \\ &= \bar{A}_{21} \hat{\Omega}_{L,11} \bar{A}_{21}'\end{aligned}$$

And given $\bar{A}_{21} = [0]$ this implies $\hat{\Omega}_{TB} = [0]$.

where $A^{-1}[1:p^s, 1:p] = \left[A_R^{-1}, D\right]$ with $D = A^{-1}[1:p^s, (p^s+1):p]$, $\hat{\Omega}_L = \begin{bmatrix} \hat{\Omega}_{LR} & 0 \\ 0 & 0 \end{bmatrix}$ with $\hat{\Omega}_{LR} = \hat{\Omega}_L[1:p^s, 1:p^s]$.

The same as before, $\hat{\Omega}_{LR}$ disregards the terms equal to zero in $\hat{\Omega}_L$, that is $\hat{\sigma}_{cc'}$ for $c \wedge c' > p^s$. Providing that $\hat{\sigma}_{cc'}$ is unbiased for all c and c' (regardless some of them are equal to zero) $\hat{\Omega}_{TR}$ is also unbiased.

References

Achlioptas, D., Clauset, A., Kempe, D., & Moore, C. (2009). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)*, 56(4), 21.

Ahmed, N. K., Neville, J., & Kompella, R. (2014). Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2), 7.

Bernstein, D. S. (2005). *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*, volume 41. Princeton University Press, Princeton.

Brawer, R. & Pirovino, M. (1992). The linear algebra of the pascal matrix. *Linear Algebra and Its Applications*, 174, 13–23.

Chandrasekhar, A. & Lewis, R. (2011). Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29(4) (pp. 251–262).: ACM.

Frank, O. (1977). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4(1), 81–89.

Frank, O. (1980). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1), 45–50.

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

Fuller, W. A. (2011). *Sampling Statistics*, volume 560. John Wiley & Sons.

Granovetter, M. (1976). Network sampling: Some first steps. *American Journal of Sociology*, 81(6), 1287–1303.

Handcock, M. S. & Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of Applied Statistics*, 4(1), 5–25.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260), 663–685.

Jackson, M. O. (2008). *Social and Economic Networks*, volume 3. Princeton University Press, Princeton.

Jackson, M. O., Rodriguez-Barraquer, T., & Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5), 1857–1897.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Science & Business Media.

Lakhina, A., Byers, J. W., Crovella, M., & Xie, P. (2003). Sampling biases in ip topology measurements. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1 (pp. 332–341).: IEEE.

Ribeiro, B. & Towsley, D. (2012). On the estimation accuracy of degree distributions from graph sampling. In *51st IEEE Conference on Decision and Control (CDC)* (pp. 5240–5247).: IEEE.

Rothenberg, R. (1995). Commentary: Sampling in social networks. *Connections*, 18, 104–110.

Santos, P. & Barrett, C. (2008). What do we learn about social networks when we only sample individuals. *Not Much. Department of Applied Economics and Management Working Paper Cornell University*.

Stumpf, M. P. & Wiuf, C. (2005). Sampling properties of random graphs: the degree distribution. *Physical Review E*, 72(3), 036118.

Thompson, S. K. (2006). Adaptive web sampling. *Biometrics*, 62(4), 1224–1234.

Yates, F. & Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2), 253–261.

Zhang, Y., Kolaczyk, E. D., Spencer, B. D., et al. (2015). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Annals of Applied Statistics*, 9(1), 166–199.

Table 1: Populations and samples, summary

Population	Definition	Example	Parameter
$U = \{u_1, u_2, \ldots, u_N\}$	Population of individual elements.	Population of houses.	-
$\Theta = \left\{\theta^1, \ldots, \theta^J\right\}$	Population of sets in which the elements of U are grouped.	Population of blocks.	T_k =Total number of sets of size k in Θ.
$\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$	Population of all possible subsets we can get from each element of Θ. It is an "artificial" population.	All possible combinations of houses we can get from each block.	L_c = Total number of subsets of size c in Γ.
$S = \{u_1, u_2, \ldots, u_n\}$	Sample of U.	Sample of houses.	-
$\Gamma^s = \{\gamma \vert \forall \gamma \in S\}$	Sample of Γ built from S.	Sample of houses grouped as they are in Θ and all possible combinations of such groups.	-

Table 2: Induced subgraph sampling $N = 300$.

		CHT estimator			Naïve estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_{\mathcal{G}}(0) = 0.120$	0.000	0.011	0.011	0.005	0.011
	$P_{\mathcal{G}}(1) = 0.163$	0.000	0.021	0.022	0.023	0.027
	$P_{\mathcal{G}}(2) = 0.293$	0.000	0.030	0.031	-0.010	0.021
	$P_{\mathcal{G}}(3) = 0.283$	-0.001	0.036	0.035	-0.061	0.064
	$P_{\mathcal{G}}(4) = 0.140$	0.001	0.025	0.026	-0.056	0.058
	Average degree 2.16	0.000	0.047	0.049	-0.213	0.219
0.7	$P_{\mathcal{G}}(0) = 0.120$	0.000	0.030	0.029	0.022	0.027
	$P_{\mathcal{G}}(1) = 0.163$	0.000	0.074	0.071	0.047	0.052
	$P_{\mathcal{G}}(2) = 0.293$	-0.007	0.118	0.111	-0.079	0.082
	$P_{\mathcal{G}}(3) = 0.283$	0.006	0.125	0.112	-0.173	0.174
	$P_{\mathcal{G}}(4) = 0.140$	0.000	0.068	0.062	-0.116	0.117
	Average degree 2.16	0.008	0.104	0.103	-0.643	0.647
0.5	$P_{\mathcal{G}}(0) = 0.120$	0.005	0.080	0.079	0.040	0.044
	$P_{\mathcal{G}}(1) = 0.163$	-0.017	0.260	0.246	0.021	0.029
	$P_{\mathcal{G}}(2) = 0.293$	0.032	0.440	0.388	-0.176	0.178
	$P_{\mathcal{G}}(3) = 0.283$	-0.033	0.417	0.331	-0.248	0.249
	$P_{\mathcal{G}}(4) = 0.140$	0.013	0.118	0.123	-0.135	0.135
	Average degree 2.16	0.000	0.183	0.175	-1.082	1.087

Table 3: Induced subgraph sampling $N = 600$.

		CHT estimator			Naïve estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_{\mathcal{G}}(0) = 0.131$	0.000	0.007	0.007	0.001	0.007
	$P_{\mathcal{G}}(1) = 0.136$	0.000	0.015	0.015	0.029	0.031
	$P_{\mathcal{G}}(2) = 0.286$	-0.002	0.024	0.025	0.000	0.015
	$P_{\mathcal{G}}(3) = 0.336$	0.001	0.021	0.023	-0.086	0.087
	$P_{\mathcal{G}}(4) = 0.108$	0.000	0.015	0.015	-0.044	0.045
	Average degree 2.15	0.000	0.034	0.035	-0.212	0.217
0.7	$P_{\mathcal{G}}(0) = 0.131$	0.000	0.020	0.020	0.013	0.017
	$P_{\mathcal{G}}(1) = 0.136$	0.000	0.052	0.052	0.065	0.067
	$P_{\mathcal{G}}(2) = 0.286$	0.002	0.085	0.083	-0.063	0.066
	$P_{\mathcal{G}}(3) = 0.336$	-0.002	0.073	0.073	-0.224	0.225
	$P_{\mathcal{G}}(4) = 0.108$	0.000	0.039	0.040	-0.089	0.090
	Average degree 2.15	0.000	0.069	0.073	-0.645	0.649
0.5	$P_{\mathcal{G}}(0) = 0.131$	0.000	0.058	0.056	0.028	0.031
	$P_{\mathcal{G}}(1) = 0.136$	0.000	0.179	0.174	0.046	0.049
	$P_{\mathcal{G}}(2) = 0.286$	0.010	0.272	0.260	-0.167	0.168
	$P_{\mathcal{G}}(3) = 0.336$	-0.011	0.221	0.203	-0.301	0.302
	$P_{\mathcal{G}}(4) = 0.108$	0.003	0.098	0.079	-0.104	0.104
	Average degree 2.15	0.000	0.128	0.124	-1.077	1.081

Table 4: Incident subgraph sampling $N = 300$.

		CHT estimator			Naïve estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_{\mathcal{G}}(0) = 0.120$	0.000	0.008	0.008	0.020	0.021
	$P_{\mathcal{G}}(1) = 0.163$	0.000	0.018	0.018	0.044	0.046
	$P_{\mathcal{G}}(2) = 0.293$	0.000	0.027	0.027	0.019	0.027
	$P_{\mathcal{G}}(3) = 0.283$	0.000	0.025	0.026	-0.035	0.039
	$P_{\mathcal{G}}(4) = 0.140$	0.000	0.012	0.013	-0.047	0.048
0.7	$P_{\mathcal{G}}(0) = 0.120$	-0.001	0.022	0.022	0.083	0.084
	$P_{\mathcal{G}}(1) = 0.163$	0.003	0.054	0.057	0.139	0.141
	$P_{\mathcal{G}}(2) = 0.293$	-0.001	0.085	0.085	0.011	0.024
	$P_{\mathcal{G}}(3) = 0.283$	-0.001	0.083	0.078	-0.127	0.129
	$P_{\mathcal{G}}(4) = 0.140$	0.000	0.036	0.035	-0.106	0.106
0.5	$P_{\mathcal{G}}(0) = 0.120$	-0.001	0.052	0.054	0.198	0.199
	$P_{\mathcal{G}}(1) = 0.163$	0.007	0.160	0.164	0.206	0.208
	$P_{\mathcal{G}}(2) = 0.293$	-0.012	0.242	0.250	-0.061	0.065
	$P_{\mathcal{G}}(3) = 0.283$	0.008	0.199	0.204	-0.212	0.212
	$P_{\mathcal{G}}(4) = 0.140$	-0.002	0.078	0.076	-0.131	0.131

Table 5: Incident subgraph sampling $N = 600$.

		CHT estimator			Naive estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_{\mathcal{G}}(0) = 0.131$	0.000	0.005	0.005	0.017	0.017
	$P_{\mathcal{G}}(1) = 0.136$	0.000	0.011	0.012	0.047	0.048
	$P_{\mathcal{G}}(2) = 0.286$	0.000	0.018	0.019	0.032	0.034
	$P_{\mathcal{G}}(3) = 0.336$	0.000	0.018	0.018	-0.059	0.060
	$P_{\mathcal{G}}(4) = 0.108$	0.000	0.008	0.008	-0.037	0.037
0.7	$P_{\mathcal{G}}(0) = 0.131$	0.000	0.014	0.015	0.076	0.076
	$P_{\mathcal{G}}(1) = 0.136$	0.001	0.039	0.040	0.151	0.152
	$P_{\mathcal{G}}(2) = 0.286$	-0.001	0.059	0.061	0.031	0.034
	$P_{\mathcal{G}}(3) = 0.336$	0.002	0.052	0.051	-0.176	0.176
	$P_{\mathcal{G}}(4) = 0.108$	-0.001	0.023	0.022	-0.082	0.082
0.5	$P_{\mathcal{G}}(0) = 0.131$	-0.002	0.035	0.037	0.188	0.189
	$P_{\mathcal{G}}(1) = 0.136$	0.009	0.109	0.115	0.227	0.229
	$P_{\mathcal{G}}(2) = 0.286$	-0.017	0.165	0.170	-0.049	0.050
	$P_{\mathcal{G}}(3) = 0.336$	0.012	0.137	0.131	-0.266	0.266
	$P_{\mathcal{G}}(4) = 0.108$	0.003	0.052	0.048	-0.101	0.101