On Set Size Distribution Estimation and the Characterization of Large Networks via Sampling

Fabricio Murai, Bruno Ribeiro, Don Towsley, and Pinghui Wang

Abstract—In this work we study the set size distribution estimation problem, where elements are randomly sampled from a collection of non-overlapping sets and we seek to recover the original set size distribution from the samples. This problem has applications to capacity planning and network theory. Examples of real-world applications include characterizing in-degree distributions in large graphs and uncovering TCP/IP flow size distributions on the Internet. We demonstrate that it is difficult to estimate the original set size distribution. The recoverability of original set size distributions presents a sharp threshold with respect to the fraction of elements that remain in the sets. If this fraction lies below the threshold, typically half of the elements in power-law and heavier-than-exponential-tailed distributions, then the original set size distribution is unrecoverable. We also discuss practical implications of our findings.

Index Terms—Cramér-Rao lower bound, Fisher information, set size distribution estimation.

I. INTRODUCTION

Networks are increasingly large and complex, posing tremendous challenges to their characterization in the wild. Characterizing network structure (e.g., degree distribution), network traffic flows (e.g., TCP/IP flow sizes in communication networks), node labels (e.g., group memberships), is usually impossible without resorting to sampling due to the size and scale of current networks. Practitioners often sample networks to estimate their characteristics. Many problems in network characterization through sampling can be mapped into the set size distribution estimation problem. The set size distribution estimation problem is stated as follows. Consider a collection of non-overlapping sets whose elements are probabilistically sampled. The problem is to estimate the underlying pre-sampling set size distribution of the data based on the samples.

Set size distribution estimation has several applications. One example of particular interest is the estimation of in-degree distributions of on-line social networks, where nodes represent people and a directed edge represents, for instance, one or more messages exchanged between two pairs of nodes. By monitoring message exchanges one samples a fraction of the edges. Using these samples we want to estimate the in-degree or out-degree distribution of nodes. The set size distribution problem also manifests itself in other areas, including Internet traffic monitoring, e.g., estimating the distribution (in packets) of TCP/UDP flow sizes [3], and in next generation Internet capacity planning, such as estimating the number of copies of a movie in a CDN of next-generation routers. Fortunately, simple maximum likelihood [3] or Bayesian-style estimators exist, even when we are unable to observe sets without observed elements.

Despite the importance of characterizing set size distributions, to the best of our knowledge no deep analysis of set size distribution estimation exists in the literature. We fill this gap and show that set size distribution estimation exhibits intriguing abnormal statistical properties. To best illustrate our results, consider the estimation of in-degree distributions of arbitrarily large power-law graphs. We prove that if less than 50% of the edges are observed then the output of any estimator (be it frequentist or Bayesian) will be as truthful to the original in-degree distribution as a set of random numbers between zero and one. Moreover, when we only observe nodes with at least on edge sampled, even a first order metric like average degree is subject to the same threshold behavior, i.e., sampling less than 50% of all incoming edges impedes the estimation of in-degree averages. We prove these and other results in the general setting of sets with arbitrary set size distributions. In what follows we give an overview of our contributions.

A. General Observations

In this work we uncover intriguing set size distribution estimation properties, including:

• A (finite) increase in samples can result in no reduction in estimation errors.

Unlike estimation problems such as election polls, where a sufficient increase in samples always results in increased accuracy, we show, paradoxically, that in the set size distribution estimation problem an increase in samples may, in practice, result in no increase in accuracy. Section IV unveils the root cause of this odd behavior and explains how to avoid it. Another interesting property is:

• In networks with power-law set size distributions (our results hold for any heavier-than-exponential distributions), randomly sampling less than 50% of the set elements (e.g., edges of a node) provides almost no information about the set size distribution or the average set size.
The other hand, accurate set size distribution estimation is always possible in networks with sub-exponential set size distributions.

The above observation is interesting because power-law distributions have more tail probability mass and, thus, large sets are more likely to have sampled elements than when the distributions have sub-exponential tails. However, and despite this, we show that if less than 50% of elements are sampled, then estimates of power-laws distributions (more precisely, any heavier-than-exponential distribution) are significantly less accurate than the estimates obtained from sub-exponential distributions. Our work also provides a host of equally puzzling observations, fully and formally presented in Section IV.

B. Outline

Our paper is organized as follows. In Section II we conduct experiments on the indegree distribution estimation with real data. Section III presents the sampling and estimation models. Section IV presents our theoretic results. Section VI presents our discussion section where we analyze problems that field analysts are likely to face in practice, highlighting common mistakes made in the literature and how to avoid them. Finally Section VII presents the conclusions and related work.

II. Estimation with Real Data

In this section, we experiment with one particular application of the set size distribution problem: the estimation of the in-degree distribution of a network. Consider the Enron email dataset [6], that describes a network composed by a group of people who exchanged emails during a certain period of time. Here each node represents a person and two people have a directed edge if one has emailed the other. The maximum in-degree in this network is 1383.

Collecting a fraction of the exchanged messages means sampling network edges. Disregarding edge weights, assume directed edges are independently sampled with probability $p$. Henceforth, each person with more than one observed incoming email will be called a sample. Figure 1a depicts the quality of the in-degree estimator in (4) (see Section IV for the derivation) with $p = 0.25$, leading to $N = 10^4$ sampled individuals. The black dots indicate the true in-degree distribution, the blue curve shows a typical estimate, and the heat map indicates the density of estimated values across 100 runs, where red indicates high density and yellow (white) indicates low (no) density of estimated values. We observe from the blue curve that the estimated values can be orders of magnitude away from the actual values and from the heat map we observe that the blue line is typical.

In what follows we illustrate the effects of varying the number of samples $N$ or changing the sample probability $p$ separately. To vary $N$ while keeping $p$ fixed, we draw a node in-degree directly from the in-degree distribution of this network and subsequently sample its edges. We repeat this process until we obtain $N$ observed sets. This can be seen as sampling a larger (smaller) network that has the same degree distribution.

We make two main observations:

1) Increasing the number of samples does not reduce estimation error. This is an odd behavior. We know from estimation theory that the error should decrease by $\sqrt{M}$ when the number of samples is increased by a factor of $M$. Figure 1b shows the corresponding results for $N = 50 \times 10^3$. We observe that the estimated fraction of nodes of each degree can still be very far from the actual values.

To make it clear that the accuracy gain from increasing the number of samples is not in agreement with theory, we compute the estimate error obtained when we vary the number of samples $N \in \{5, 10, 20, 50, 100\} \times 10^3$, for $p = 0.25$. The error is first measured in terms of the Normalized Root Mean Square Error (NRMSE), which is defined as

$$\text{NRMSE}(\hat{\theta}_i) = \frac{\sqrt{E[(\hat{\theta}_i - \theta_i)^2]}}{\theta_i}.$$  

where $\hat{\theta}_i$ and $\theta_i$ are the estimated and true fraction of degree $i$ nodes, respectively. Then we take the average NRMSE from the head (degrees up to 10) and the tail (degrees larger than 10) of the distribution separately. Surprisingly, we observe in Figure 1c that there is almost no improvement in accuracy across different sample sizes, even when we compare $5 \times 10^3$ and $10^5$ samples. We also display in this figure the expected reduction in the NRMSE for both head and tail by dashed lines. It turns out that the error does not decrease as we would expect. This raises the question of why, which we address in Section IV.

2) For much larger values of $p$, the error starts to decrease with the number of samples. According to Theorem 4.1 presented in Section IV, the difficulties experienced above arise due to the use of small sampling probability ($p < 0.5$) with heavy-tailed distributions, and not due to a lack of samples. Hence we repeat the experiment using $p = 0.9$. Figures 1d and 1e show the heat maps for $N = 20 \times 10^3$ and $N = 10^5$. As opposed to what we previously saw, increasing the number of samples does increase the accuracy of the estimate. The accuracy gain as a function of the number of samples is shown in Figure 1f. In fact, we observe that the NRMSE does decrease as expected for the head of the distribution, but not for the tail. Why are there two distinct behaviors, one for the head and one for the tail? Why did it help to increase the number of samples when estimating frequencies of small degrees for $p = 0.9$, as opposed to what we observed for $p = 0.25$? Is it possible to make the NRMSE of the tail to decrease as fast as the NRMSE of the head?

In order to investigate the questions we pose here, we study the Cramér-Rao Lower Bound (CRLB) of the set size estimation problem. This give us a lower bound on the estimation errors based on the amount of information contained in the samples, measured in terms of Fisher Information. Moreover, we apply the CRLB to the estimation of the in-degree distribution and average in-degree.
III. Model

Let \( S_k \) be a nonempty set of elements, \( k = 1, \ldots, n \), with \( S_i \cap S_j = \emptyset \), \( i, j = 1, \ldots, n \), \( i \neq j \). Let \( S_k = |S_k| \) denote the size of the \( k \)-th set and assume set sizes are i.i.d. with distribution \( S_k \sim \theta = (\theta_1, \ldots, \theta_W) \), \( W > 1 \), \( k \geq 1 \). We assume \( W \) is finite (\( W < \infty \)). The model divides elements into groups (sets) and our task is to characterize those groups from an incomplete observation (sample) of these groups. To illustrate the model, consider a directed graph; edges can be grouped by the nodes they are incident to (depart from), in which case \( S_k \) is the set of incoming (outgoing) edges of a node \( k \), \( \theta \) is the indegree (outdegree) distribution, and \( W \) is the maximum indegree (outdegree). Another straightforward example is characterizing IP traffic in a communications network, where \( k \) is a TCP flow, \( S_k \) is the set of TCP/IP packets that constitute flow \( k \), and \( W \) is the maximum observable flow size.

Sampling

We observe (sample) elements of \( S_k \), \( k = 1, \ldots, n \), with probability \( p \) - a process also known as thinning. Let \( \alpha(S_k) \) be a random function that returns the number of observed elements of \( S_k \). Elements are sampled independently (i.e., the sampling process is Bernoulli) and thus,

\[
P(\alpha(S_k) = j | S_k = i) = \begin{cases} \binom{i}{j} p^j q^{i-j}, & 0 \leq j \leq i, i > 1 \\ 0, & \text{otherwise} \end{cases}
\]

where \( q = 1 - p \). We assume that when no elements of a set are observed, then the set as a whole is not observed, i.e., \( S_k \) is said to be observable only if \( \alpha(S_k) > 0 \). Thus, we denote

\[
S = \{ \alpha(S_k) : \alpha(S_k) > 0, k = 1, \ldots, n \}
\]

the set of the observable set sizes. Let \( N = |S| \) denote the number of observed sets.
**Estimation**

We start by considering $p = 1$, that is, all elements of all sets are observed. The minimum variance estimator of $\theta_i$ is

$$
T_i^*(S_1, \ldots, S_n) = \frac{1}{N} \sum_{k=1}^{n} 1\{S_k = i\},
$$

where $N = n$. To measure the accuracy of the estimates we consider the mean squared error (MSE) – a.k.a. quadratic loss – of the estimates

$$
\text{MSE}(T_i^*(S_1, \ldots, S_n)) = E[(T_i^*(S_1, \ldots, S_n) - \theta_i)^2]
$$

from $S_i(1)$ to $S_i(W)$ is constant and $\sum_{j=1}^{W} j^2\theta_j = \omega(1)$, provided $p = 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = \Theta(1/N)$, provided $\theta_p \geq (a + 1)/a$.}

The lower bounds of type $\Omega(1/N)$ in Theorem 4.1 are only meaningful if they are achievable. We investigate this achievability question in our Technical Report [8], showing that, in fact, there exists a Maximum Likelihood Estimator (MLE) $T_i^*(S)$ of $\theta_i$, $i = 1, \ldots, W$ that is asymptotically efficient and normal, which means that $T_i^*(S)$ approaches the CRLB uniformly as $N \to \infty$. Hence, the corresponding bounds for $T_i^*(S)$ are as follows.

**Theorem 4.2:** Let $\theta = (\theta_1, \ldots, \theta_W)$ be the set size distribution, $S$ be the sequence of observed set sizes after randomly sampling elements of the sets with probability $p$, and $T_i^*(S)$, $i \geq 1$ is the MLE of $\theta_i$. The lower bounds for $T_i^*(S)$ are as follows.

1. If $\theta_p$ decreases faster than exponentially in $W$, i.e., $-\log \theta_p = \omega(W)$, then $\text{MSE}(T_i(S)) = \Omega(1/N)$ provided $0 < p < 1$.
2. If $\theta_p$ decreases exponentially in $W$, i.e., $\log \theta_p = W \log a + o(W)$ as for some $0 < a < 1$, then
   a) $\log \text{MSE}(T_i(S)) = \Omega(W - \log N)$, provided $p < a/(a + 1)$,
   b) $\text{MSE}(T_i(S)) = \Omega(W^{2a+1}/N)$, provided $p = a/(a + 1)$,
   c) $\text{MSE}(T_i(S)) = \Omega(1/N)$, provided $p > a/(a + 1)$.
3. If $\theta_p$ decreases more slowly than exponential, i.e., $-\log \theta_p = o(W)$, then
   a) $\log \text{MSE}(T_i(S)) = \Omega(W - \log N)$, provided $p < 1/2$,
   b) $\text{MSE}(T_i(S)) = \omega(1/N)$, provided $p = 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = \omega(1)$,
   c) $\text{MSE}(T_i(S)) = \Omega(1/N)$, provided either $p > 1/2$ or $p < 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = O(1)$.

In what follows we focus on unbiased estimates; our results can be applied on the following scenarios: (1) flow distribution estimation in IP traffic and (2) latent in-degree distribution estimation on directed graphs (e.g. Web graph). In the context of graphs, sampling edges with probability $p$ creates a dependence between $N$, $W$, and $\theta$, that we address in Section VI-C.

**Theorem 4.1:** Let $\theta = (\theta_1, \ldots, \theta_W)$ be the set size distribution, $S$ be the sequence of observed set sizes after randomly sampling elements of the sets with probability $p$, and $T_i(S)$, $i \geq 1$ be an unbiased estimator of $\theta_i$.

1. If $\theta_p$ decreases faster than exponentially in $W$, i.e., $-\log \theta_p = \omega(W)$, then $\text{MSE}(T_i(S)) = \Omega(1/N)$ provided $0 < p < 1$.
2. If $\theta_p$ decreases exponentially in $W$, i.e., $\log \theta_p = W \log a + o(W)$ as for some $0 < a < 1$, then
   a) $\log \text{MSE}(T_i(S)) = \Omega(W - \log N)$, provided $p < a/(a + 1)$,
   b) $\text{MSE}(T_i(S)) = \Omega(W^{2a+1}/N)$, provided $p = a/(a + 1)$,
   c) $\text{MSE}(T_i(S)) = \Omega(1/N)$, provided $p > a/(a + 1)$.
3. If $\theta_p$ decreases more slowly than exponential, i.e., $-\log \theta_p = o(W)$, then
   a) $\log \text{MSE}(T_i(S)) = \Omega(W - \log N)$, provided $p < 1/2$,
   b) $\text{MSE}(T_i(S)) = \omega(1/N)$, provided $p = 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = \omega(1)$,
   c) $\text{MSE}(T_i(S)) = \Omega(1/N)$, provided either $p > 1/2$ or $p < 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = O(1)$.

In what follows we consider the problem of estimating the average set size $m_\theta = \sum_{i=1}^{W} i\theta_i$ from the sample $S$. Surprisingly, we obtain bounds analogous to the bounds for the set size distribution in Theorem 4.1. This result is surprising because the average observed set size $m_\phi = \sum_{i=1}^{W} i\phi_i$ have remarkably different bounds: $m_\phi$ is always finite (independent of $p$ or $W$) as long as the second moment of $\phi$ is finite (see Section V).

**Theorem 4.3:** Let $\theta = (\theta_1, \ldots, \theta_W)$ be the set size distribution, $S$ be the sequence of observed set sizes after randomly sampling elements of the sets with probability $p$, and $\hat{m}_\theta(S)$ be an unbiased estimator of $m_\theta$.

1. If $\theta_p$ decreases faster than exponentially in $W$, i.e., $-\log \theta_p = \omega(W)$, then $\text{MSE}(\hat{m}_\theta(S)) = \Omega(1/N)$ provided $0 < p < 1$.
2. If $\theta_p$ decreases exponentially in $W$, i.e., $\log \theta_p = W \log a + o(W)$ for some $0 < a < 1$, then

IV. RESULTS

In this section we introduce our main results which derive MSE lower bounds for unbiased set size distribution and average set size estimators. In addition, we show that the lower bounds obtained for the set size distribution are achievable by a Maximum Likelihood Estimator. We consider a general formulation of the sampling problem, where the number of observed sets $N$ is constant and $N$ is independent of the maximum degree $W$. We also consider the sampling probability $p$ to be a known constant. Later in Section VI we discuss the validity of these assumptions on real world applications, showing how our results can be applied on the following scenarios: (1) flow size distribution estimation in IP traffic and (2) latent in-degree distribution estimation on directed graphs (e.g. Web graph).

In the context of graphs, sampling edges with probability $p$ creates a dependence between $N$, $W$, and $\theta$, that we address in Section VI-C.
where

\[ S \]

the smallest MSE that any unbiased estimator

\[ d \]

and

\[ \text{if} \]

\[ S \]

in our model, note that for a random observed set

\[ T \]

A. Lower Bound on Estimation Errors

In this section we derive a lower bound on the Mean Squared Error (MSE) of \( T_i(\mathbf{S}) \), \( i = 1, \ldots, W \). For this we use the Cramér-Rao (CR) lower bound of \( T_i(\mathbf{S}) \), which gives the smallest MSE that any unbiased estimator \( T_i \) can achieve.

Recall that a set is observable only if one or more of its elements are observable. The probability that a (random) set \( S \) has \( k \) observed elements, given \( j > 0 \), is defined as

\[ b_{ji}(p) = P[\alpha(S) = j | \alpha(S) > 0, |S| = i] = \frac{(i)^{p}q^{i-j}}{1 - q^i}, \quad (1) \]

if \( 0 < j \leq i \leq W \) and \( b_{ji}(p) = 0 \) otherwise, where \( q = 1 - p \) and \( \alpha(S) > 0 \) is the size of \( S \) after sampling. Let \( d_j(\theta, p) \) denote the fraction of observed sets with exactly \( j \) observed elements. From (1) we have, \( j = 1, \ldots, W \),

\[ d_j(\theta, p) = P[\alpha(S) = j | |S| > 0] = \sum_{i=j}^{W} P[\alpha(S) = j | \alpha(S) > 0, |S| = i] \times P[|S| = i | \alpha(S) > 0] = \sum_{i=j}^{W} b_{ji}(p) \phi_i(\theta). \quad (2) \]

where

\[ \phi_i(\theta) = P[|S| = i | \alpha(S) > 0] = \frac{\theta_i(1 - q^i)}{\sum_{k=1}^{W} \theta_k(1 - q^k)}, \quad (3) \]

is the distribution observed set sizes. Or, in matrix notation,

\[ d(\theta, p) = (d_1(\theta, p), \ldots, d_W(\theta, p))^T \]

and \( B(p) = [b_{ji}(p)] \), \( j = 1, \ldots, W. \) To illustrate the distribution \( d(\theta, p) \) in our model, note that for a random observed set \( S \),

\[ \alpha(S) \sim d(\theta, p), \]

with likelihood function

\[ P[\alpha(S) = j | \theta] = (B(p)\phi(\theta))_j = d_j(\theta, p), \quad j = 1, \ldots, W. \quad (4) \]

In what follows for simplicity we denote \( d_j(\theta, p) \) by \( d_j(\theta) \).

We now apply the Cramér-Rao Theorem to find the lower bounds of MSE(\( T_i(\mathbf{S}) \)). The Cramér-Rao Theorem states that the MSE of any unbiased estimator \( T \) is lower bounded by the inverse of the Fisher information matrix divided by the number of independent samples \( N \), provided some weak regularity conditions hold [11, Chapter 2], i.e.,

\[ \text{MSE}(T_i(\mathbf{S})) = E[(T_i(\mathbf{S}) - \theta_i)^2] \geq \frac{(J(\theta)(p))^{-1}}{N}, \quad 1 \leq i \leq W. \quad (5) \]

where \( (J(\theta)(p))^{-1} \) is the inverse of the Fisher information matrix of a single set size observation defined using the likelihood function (4) as

\[ (J(\theta)(p))_{i,k} = \sum_{j=1}^{W} \frac{\partial d_j(\phi(\theta))}{\partial \theta_i} \frac{\partial d_j(\phi(\theta))}{\partial \theta_k} \frac{1}{d_j(\phi(\theta))}, \quad (6) \]

given \( \sum_{i=1}^{W} \phi_i = 1 \).

The lower bound in (5) is known in the literature as the Cramér-Rao lower bound or CRLB for short. Let \( T_i(\mathbf{S}) \) be an unbiased estimator, \( i = 1, \ldots, W. \) We say \( T_i(\mathbf{S}) \) is asymptotically efficient if MSE(\( T_i(\mathbf{S}) \)) approaches the Cramér-Rao lower bound in (5) as \( N \to \infty \). We show in [8], that the Maximum Likelihood Estimator is asymptotically efficient on the set size estimation under the condition that the bound is finite. In what follows we represent \( J(\theta)(p) \) as \( J(\theta) \) for simplicity.

B. Obtaining the CRLB

In what follows we derive the CRLB in closed-form as a function of the original set size distribution \( \theta \), the sampling probability \( p \), and the number of observed sets \( N \), where we ignore the constraint \( \sum_{i=1}^{W} \theta_i = 1 \). Deriving a closed-form solution for the inverse of \( J(\theta) \) is no easy task as matrix \( J(\theta) \) is a function of \( \partial \phi(\theta)/\partial \theta_i \), \( j = 1, \ldots, W. \), which makes \( J(\theta) \) a non-linear function of \( \theta \). The Fisher information matrix in (6) can be derived as a function of \( \phi \) and thus

\[ J_{i,k}(\phi) = \sum_{j=1}^{W} \frac{\partial d_j(\phi)}{\partial \phi_i} \frac{\partial d_j(\phi)}{\partial \phi_k} \frac{1}{d_j(\phi)}, \quad (7) \]

given \( \sum_{i=1}^{W} \phi_i = 1 \); and because \( d_j(\phi) \) is linear in \( \phi \) the above yields

\[ (J(\phi))^{-1} = B(p)^{-1} \text{diag}(B(p)\phi)^{-1}(B(p)^{-1})^T - \phi \phi^T. \quad (8) \]

Here the term \( \phi \phi^T \) corresponds to the accuracy gain obtained by considering the constraint \( \sum_{i=1}^{W} \phi_i = 1 \) (see Tune and Veitch [10] for more details and Gorman and Hero [4] for the general formula on adding equality constraints to the CRLB). Quantitatively we can safely ignore the constant term \( \phi \phi^T \) as we are interested in the behavior of \( (J(\phi))^{-1} \) as a function of \( W \) and the elements of \( \phi \phi^T \) must be smaller than one. All that is left to do is to find a relationship between \( (J(\phi))^{-1} \) and \( (J(\theta))^{-1} \).
We now obtain \((J^{(\theta)})^{-1}\) from \((J(\phi))^{-1}\) through the Jacobian \(\nabla H = [h_{ik}], h_{ik} = \partial \theta_i(\phi)/\partial \phi_k\) with \(\theta_i(\phi)\) obtained from inversion of (3), we arrive at the equivalent multivariate rule [11, pp. 83] and express \((J(\theta))^{-1}\) as
\[
(J^{(\theta)})^{-1} = \nabla H (J(\phi))^{-1} \nabla H^\top.
\] (9)

Using (8) – detailed derivation relegated to the Technical Report [8] – we find:
\[
[(J^{(\theta)})^{-1}]_{ij} = \sum_{k=\max(i,j)}^{W} \binom{q}{p}^{2k} \binom{k}{j} \binom{k}{i} (-1)^{-i-j} (q^i - 1) 
\times (q^{-j} - 1) d_j(\theta).
\] (10)

Substituting (10) into (9) – and applying a variety of algebraic manipulations detailed in the Technical Report – yields
\[
[(J^{(\theta)})^{-1}]_{ii} = \frac{1}{\eta^2} \left( \frac{1}{(1 - q^2)^2} [(J^{(\theta)})^{-1}]_{ii} \right)
+ \theta_i^2 \sum_{j=1}^{W} \sum_{k=1}^{W} \binom{k}{j} (1 - q^i) (1 - q^{i'})
- 2 \theta_i \sum_{j=1}^{W} \binom{k}{j} (1 - q^i) (1 - q^{i'})
\]
where \(\eta = \sum_{j=1}^{W} \phi_j(\theta)/(1 - q^2)\). Note that term \(A_1(i)\) is proportional to the CRLB of \(\phi_i[(J^{(\theta)})^{-1}]_{ii}\) but terms \(A_2(i)\) and \(A_3(i)\) are more involved. Through a series of algebraic manipulations of terms \(A_1, A_2,\) and \(A_3\), all detailed in the Technical Report, we find that \(A_1(i) + A_2(i) - A_3(i)\) grows as a function of \((1 - p)/p\) and \(W\), yielding the relation
\[
\text{MSE}(T_i(S)) = \Omega \left( \frac{\sum_{j=1}^{W} (1 - p)^j \theta_j}{N} \right), \quad i = 1, \ldots, W,
\] (12)
where the number of observed sets \(N\) is large but constant with respect to \(W\).

The result in (12) is very powerful as it gives a simple estimation error lower bound as a function of the sampling probability \(p\) and the original set size distribution \(\theta\). In particular, the following examples applied to (12) give some intuition on the results in Theorem 4.1 – a detailed exposition is presented in the Technical Report. For instance, a close look at (12) reveals that when \(((1 - p)/p)^{\theta_i} = \Omega(i^{-1})\) for all \(i > i^*, i^* \ll W\), then the sum in (12) grows at least as fast as the harmonic series, which grows as \(\log W\). On the other hand, when \(((1 - p)/p)^{\theta_i} = O(i^{-\beta}), \beta > 1\), the sum in (12) converges to a constant, more precisely, it grows no faster than a Riemann zeta function with parameter \(\beta, \zeta(\beta)\).

Thus, for a given \(\theta\) with \(W \gg 1\) the CRLB exhibits an interesting sharp threshold \((p_0)\) related to the sampling probability \(p\). If \(p < p_0\) no estimator \(T_i(\theta), i = 1, \ldots, W\), is able to achieve accurate estimates of \(\theta_i\). If \(p > p_0\), there exist estimators \(T_i(S), i = 1, \ldots, W\) that can achieve accurate estimates, as \(N \to \infty\). To be more specific, we look at the threshold behavior of \(p\) by breaking down \(\theta\) into three broad classes of distributions:

1) If \(\theta_W\) decreases faster than exponentially in \(W\) there is no threshold behavior of \(p\). This is because when \(- \log \theta_W = \omega(W)\) there exists a constant \(a < 1\) such that \(((1 - p)/p) \theta_j < a^j, j = 1, 2, \ldots\). Hence, the sum in (12) converges to a constant for any \(p > 0\), yielding \(\text{MSE}(T_i(S)) = \Omega(1/N), 0 < p < 1\). Detailed arguments are presented in the Technical Report.

2) If \(\theta_W = W \log a + o(W)\) and \(p \leq a/(a + 1)\), then \(((1 - p)/p)^{\theta_j} = a^{-j} \theta_j = \Omega(1), \forall j\). Hence, the sum in (12) diverges with \(W\). On the other hand, if \(p > a/(a + 1)\) the sum in (12) converges to a constant. Detailed arguments are presented in the Technical Report.

3) Finally, if \(\theta_W\) decreases more slowly than exponential and \(p < 1/2\), then there exists an \(\epsilon > 0\), such that \(((1 - p)/p)^{\theta_j} = O((1 + \epsilon/2)^j), \forall j\). Because \(\theta_j\) decreases more slowly than an exponential the sum in (12) diverges with \(W\). If \(p \geq 1/2\) the lower bound in (12) converges to a constant. Detailed arguments are presented in the Technical Report.

To illustrate our results, we compute the MSE lower bounds in (11) where \(\theta\) is the in-degree distribution of the Enron email dataset truncated at different values of \(W\). More precisely, we take the in-degree distribution of the Enron dataset (discussed in Section II) and truncate the maximum degree to \(W\) by accumulating in \(W\) all the probability mass previously corresponding to degrees greater than \(W\). The Enron in-degree distribution is a (truncated) heavier-than-exponential distribution.

Figures 2a and 2b show the MSE lower bounds for \(p \in \{0.25, 0.90\}\), respectively. We observe that for \(p = 0.25\) (Figure 2(a)) the MSE lower bound grows with \(W\) even for small degrees, as predicted by Theorem 4.1. While, for \(p = 0.9\) (Figure 2(b)) the MSE lower bound behaves (mostly) independent of \(W\), also as predicted by Theorem 4.1. These results corroborate to explain the simulations results in Section II.

Other metrics besides the set size distribution are of interest. In what follows we observe that the accuracy is similar to that of set size distribution estimators \(T_i, i = 1, \ldots, W\). We then analyze the accuracy of the average set sizes.
V. Accuracy of Estimated Averages

In this section we consider the accuracy of unbiased average set size estimates. Let \( m_\theta = \sum_{j=1}^W j \theta_j \) be the average set size. Theorem 4.3 implies that estimating the average set size is in the same order of hardness as estimating the entire set size distribution (see proof in Technical Report). However, we show that the average size of the observed sets, i.e., the average set size in respect to \( \phi \), \( m_\phi = \sum_{j=1}^W j \phi_j \), is much easier to estimate.

Theorem 4.3 shows that estimating the average set size is asymptotically as hard as estimating the distribution \( \theta \). However, the average size of the observed sets, i.e., the average set size in respect to \( \phi \),

\[
m_\phi = \sum_{j=1}^W j \phi_j,
\]

is much easier to estimate accurately. A estimation error bound for \( m_\phi \) is affected only by the first and second moments of \( \phi \), that is, as long as \( m_\phi \) and

\[
m_\phi^{(2)} = \sum_{j=1}^W j^2 \phi_j
\]

are finite, \( m_\phi \) can be accurately estimated if enough sets are observed. To prove this, let \( \hat{m}_\phi(S) \) denote an unbiased estimate of \( m_\phi \) and let

\[
\text{MSE}(\hat{m}_\phi(S)) = E[(\hat{m}_\phi(S) - m_\phi)^2]
\]

denote the MSE of \( \hat{m}_\phi(S) \). After applying a variety of algebraic manipulations detailed in the Technical Report we arrive at the following inequality

\[
\text{MSE}(\hat{m}_\phi(S)) \geq \frac{(1, \ldots, W)(J(\phi))^{-1}(1, \ldots, W)^T - m_\phi^2}{N}
\]

\[
= \sum_{k=1}^W \sum_{i=1}^k ij \left( \begin{array}{c} k \\ i \end{array} \right) \left( \begin{array}{c} k \\ j \end{array} \right) \frac{(-q)^{2k-i-j}}{p^{2k}} (1 - q^i)(1 - q^j)d_k(\phi)
\]

\[
= \left( \sum_{i=1}^W i(p^i + q^{i+1} - 2q^i + q)\phi_i \right) - m_\phi^2 / N.
\]

Interestingly,

\[
\hat{m}_\phi^*(S) = \frac{\sum_{s \in S} s}{Np} + \left( 1 - \frac{1}{p} \right) \frac{\sum_{s \in S} \mathbf{1}_{s=1}}{N}, \quad (13)
\]

is an unbiased efficient (minimum variance) estimator of \( m_\phi \), yielding

\[
\text{MSE}(\hat{m}_\phi^*(S)) = \left( \sum_{i=1}^W i(p^i + q^{i+1} - 2q^i + q)\phi_i \right) - m_\phi^2 / N.
\]

Alternatively we can rewrite the above as

\[
\text{MSE}(\hat{m}_\phi^*) = O \left( \frac{m_\phi^{(2)} - m_\phi^2}{N} \right).
\]

Hence, \( \text{MSE}(\hat{m}_\phi^*) \) is lower bounded by the variance of the observed set sizes. A simple explanation for this behavior is likely found in the inspection paradox. The sampling is biased towards sets with larger sizes, which increases the variance of the observed set sizes and, in turn, makes it harder to unbiash the samples.

VI. Discussion

This section considers applications of our results to the estimation of Internet flow sizes on high-speed routers and to the estimation of latent indegrees (outdegrees) in directed networks (e.g. Web graph) through edge sampling, and bounds on Bayesian and biased estimators that complements our bounds of unbiased estimators. Moreover, we also discuss the practical aspects of the initialization of estimation procedures.

A. Application Example.

An important problem in network traffic measurement and planning is to estimate the distribution of the flow sizes traversing a network. In this context, packets can be seen as elements grouped by flows. Due to efficiency requirements, packets traversing a link are sampled independently with a given probability, rather than collecting information about all packets. In the most common sampling design, two parameters are chosen: \( p \), the sampling probability and \( N \), the number of flows to be observed. A router is then set to sample packets traversing a link until \( N \) different flows are sampled, but it only actually stops after the last flow terminates. Here the number of observed sets (flows) is a constant defined a priori. More generally, this sampling design can be used in any scenario where a stream of elements is to be sampled. In the above scenario \( N \) and \( W \) are fixed and our results can be directly applied. There are alternative sampling scenarios, however, where \( N \) is a random variable.

B. Variable Number of Observed Sets (\( N \)).

Our results assume that the number of observed sets \( N \), the maximum set size \( W \), and the sampling probability \( p \) are independent constants. In what follows we consider \( N \) to be a random variable that depends on the number of sets \( M \) and on the constants \( W \) and \( p \), showing how our results can still be applied.

To exemplify this scenario, consider again the Internet flow size distribution estimation, where we sample packets with probability \( \rho \) from a fixed number of flows \( M \). In this case, \( N \) is binomially distributed random variable with parameters \( M \) and \( \rho = \sum_{i=1}^W \theta_i(1 - q^i) \) (probability of observing one set chosen uniformly at random). The Chernoff bound shows that \( N \) is concentrated around its mean \( M\rho \):

\[
P(N > (1 + \delta)M\rho) \leq \exp \left( -\frac{M\rho\delta^2}{3} \right), \quad 0 < \delta < 1.
\]

For instance, choosing \( \delta = \sqrt{\frac{M}{N^2} \log \frac{1}{\epsilon}} \), yields \( P(N > (1 + \delta)M\rho) \leq \epsilon \). Even though \( M \) and \( \rho \) may be unknown in practice, this inequality still illustrates that, for a fixed \( \epsilon \), the upper bound on \( N \) increases linearly with \( M \). From this fact, it follows that the negative implications (items 2(a) and 3(a)) of Theorems 4.1, 4.2, and 4.3 hold as long as \( W \) grows faster than \( \log(M) \).
C. The Maximum Set Size $W$ as a Function of the Number of Sets $M$

Consider the following network sampling application. We wish to sample nodes on a directed network by randomly sampling edges (e.g. observing Web pages through sampling incoming links). Here a relationship between $M$ and $W$ arises, as the size of the network also determines the maximum node degree. To estimate the indegree distribution from these sampled edges we need to consider that $M$ and $W$ are coupled: the maximum indegree cannot exceed the number of nodes in the network. Nevertheless, on power-law networks, one can show that if $W$ grows as $\Omega(M^b)$ with high probability, for any $b > 0$, then we can readily apply items 2(a) and 3(a) of Theorems 4.1 and 4.3. This is true, for instance, for Barabási-Albert networks, where $W$ grows as $\Omega(M^{1/2})$ with high probability $[7]$. Since $-\log(\theta_W) \sim \log(W) = o(W)$, and clearly the number of observed sets $N \leq M$, it follows from 3(a) that when $p < 1/2$, $\log[MSE(T_i(S))] = \Omega(M^{1/2} - \log(M)) \to \infty$ as $M \to \infty$.

Also it is worth noting that all results listed in Theorems 4.1, 4.2, and 4.3 that do not depend on $W$ also hold true, even when $N$ and $W$ depend on each other. For instance, for Erdős-Rényi large networks where the degree distribution asymptotically approaches either Poisson or Normal distributions, $\theta_W$ decreases faster than exponentially and our results show that the MSE is lower bounded by $O(1/N)$ and if $N \to \infty$ then MLE achieves this bound.

D. Impact on Different Types of Estimators: Bayesian, Biased and Unbiased.

To extend our results beyond unbiased estimators we explain the connection between Fisher information, the Cramér-Rao bound and biased estimators. We also extend our results to Bayesian estimators (including maximum a posteriori estimators).

1) Extension to Biased Estimators: Let $b(\theta_i) = E[T_i(S)] - \theta_i$ be the estimator bias. Then (see Ben-Haim and Eldar [1])

$$
MSE(T_i(S)) \geq \left(1 + \frac{\partial b(\theta_i)}{\partial \theta_i}ight)^2 [(J^{(\theta)})^{-1}]_{ii},
$$

assuming $\partial b(\theta_i)/\partial \theta_i$ exists. Note if the bias derivative satisfies $-2 < \partial b(\theta_i)/\partial \theta_i < 0$, then the biased estimator has a smaller MSE than any unbiased estimator. However, we believe it is unlikely that a biased estimator can be designed to compensate for a large value of $[(J^{(\theta)})^{-1}]_{ii}$ (as large as $10^{160}$ as seen in Section IV-B for the Enron e-mail network).

2) Extension to Bayesian Estimators: Let $\theta$ now be a random variable with prior distribution $\pi_\theta$. A Bayesian estimator adds $\pi_\theta$ as extra information to the estimation problem. The Fisher information of the prior is

$$
J^{(p)}_{ij} = E \left[ \frac{\partial \ln \pi_\theta}{\partial \theta_i} \frac{\partial \ln \pi_\theta}{\partial \theta_j} \right].
$$

The Fisher information obtained exclusively by the data is $J^{(\theta)}$ presented in (6). And the total Fisher information prior + data is $[11, \text{pp. 84}]$

$$
J^{(t)} = J^{(p)} + J^{(\theta)}.
$$

The Cramér-Rao bound of a Bayesian estimator $W_i(S)$ of $\theta_i$ with prior $\pi_\theta$ yields $[11, \text{pp. 85}]$

$$
MSE(W_i(S)) \geq (J^{(t)})^{-1} = (J^{(p)} + J^{(\theta)})^{-1}.
$$

Thus, if the data contains little Fisher information then any decrease in the MSE is due to the information contained in the prior $\pi_\theta$.

E. Initialization of Estimation Procedures.

As previously stated, eq. (4) can be used to derive a maximum likelihood estimator (MLE) for $\theta$. From the MLE one could either use a constrained non-linear optimization method to maximize the likelihood function directly or use the Expectation-Maximization (EM) algorithm to write an iterative estimation procedure. In the latter case, the procedure consists of an initialization step followed by a loop of two steps known as the E-step and M-step. We discuss two issues that arise when EM is used to estimate the set size distribution.

In EM, the solution to which the algorithm converges depends on the initial guess. Therefore, in order to have an unbiased estimate, one must choose a point uniformly at random from the space of possible values. Although it may seem reasonable to choose values for each $\theta_i$ uniformly in $[0, 1]$ and then normalize them, it turns out that this does not yield uniformly distributed initial guesses. One way to correctly generate the initial guess is to sample $W - 1$ points uniformly from the unit line and then take the difference between adjacent points (including 0 and 1) $[2, \text{Chapter XI, Theorem 2.1}]$. This is equivalent to drawing from the Dirichlet distribution with $W$ parameters $\alpha = (1, \ldots, 1)$, since the Dirichlet PDF at point $\theta$ is proportional to $\prod_{i=1}^{W} \theta_i^{\alpha_i-1}$.

Nevertheless, such an initialization combined with the other two steps of EM will give us estimates $\hat{\theta}_i \in [0, 1]$ hence producing biased estimates as they are not free to assume any real value. Therefore, it is possible for the EM to achieve an MSE not in agreement with the CRLB we derived previously. This is the case when the number of samples $N$ is small and, consequently, the diagonal of $J^{-1}([\theta])$ has relatively large values (possibly greater than 1). On the other hand, for large $N$, the number of observed sets with size $i$ will converge to a Normal distribution with mean $\theta_i$ and small variance. For small enough variance, restricting $\theta_i$ to be between 0 and 1 does not affect the final estimate significantly and thus the CRLB accurately bounds the MSE.

VII. CONCLUSIONS & RELATED WORK

In this work we give explicit expressions of MSE lower bounds of unbiased estimators of the distribution of set sizes $\theta$ and the average set size $m_\theta$ with sampling probability $p$. We show that the estimation error of $\theta$ grows at least exponentially in $W$, when $\log \theta_W = W \log a + o(W)$ as $W \to \infty$ for some $0 < a < 1$, and $p > a/(a + 1)$, or when $\log \theta_W = o(W)$ as $W \to \infty$ and $p < 1/2$, which indicates that unbiased estimators of some distributions $\theta$ are too inaccurate to be useful for practitioners. Moreover we show that unbiased estimates of $m_\theta$ suffer from similar problems.
Not much prior work exists in the literature. Hohn and Veitch [5] first observed that using a sampling probability of \( p < 1/2 \) poses problems in the context of two specific estimators for the flow size distribution when the distribution obeys a power law. In particular, they showed that their estimators are asymptotically unbiased with decreasing error as the number of flow samples increases when \( p \geq 1/2 \) but not when \( p < 1/2 \). Our work shows that this is a fundamental result of set size distribution estimation and not specific to any specific estimator. Ribeiro et al. [9] was the first to introduce the use of Fisher information as a design tool for flow size estimation. Experiments reported in that paper suggested that there is little information when \( p \) is small and showed how this information can be significantly increased with the addition of other data taken from packet headers. Last, Tune and Veitch [10] applied Fisher information to compare packet sampling with flow sampling. In the process of doing so, they obtained a variety of useful Fisher information inverse identities, some of which we rely on in this work.

REFERENCES


Fabricio Murai received the B.Sc. degree in Computer Science (with high honor) and the M.Sc. degree in Computer and System Engineering both from the Federal University of Rio de Janeiro, Brazil, in 2007 and 2011, respectively. He is currently a Ph.D. student at the University of Massachusetts Amherst. His research interests include modeling, performance evaluation and network science.

Bruno Ribeiro obtained his Ph.D. degree in Computer Science from University of Massachusetts Amherst in 2010. He is currently a Post-Doctoral Researcher jointly at the University of Massachusetts Amherst and the Network Science Collaborative Technology Alliance funded by the U.S. Army Research Laboratory. Ribeiro has been a visiting researcher at Northeastern University since 2012. Ribeiro’s research focuses on applications of probability theory and statistics to complex systems, with an emphasis on the design of principled experiments to accurately characterize dynamic processes evolving on top of complex networks in the wild.

Don Towsley (M’78–SM’93–F’95) received the B.A. degree in physics and Ph.D. degree in computer science from the University of Texas at Austin in 1971 and 1975, respectively. He is currently a Distinguished Professor with the Department of Computer Science, University of Massachusetts, Amherst. His research interests include networks and performance evaluation. Prof. Towsley has been elected Fellow of the Association for Computing Machinery (ACM). He has received several ACM and IEEE best paper awards.

Pinghui Wang received the B.S. degree in information engineering and Ph.D degree in automatic control from Xi’an Jiaotong University, Xi’an, China, in 2006, 2012 respectively. From April 2012 to October 2012, he was a postdoctoral researcher with the Department of Computer Science and Engineering at The Chinese University of Hong Kong. He is currently a postdoctoral researcher with School of Computer Science at McGill University, QC, Canada. His research interests include Internet traffic measurement and modeling, traffic classification, abnormal detection, and online social network measurement. This work is done at Xi’an Jiaotong University.