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

Fabricio Murai*, Bruno Ribeiro*, Don Towsley*, and Pinghui Wang†
*Computer Science Department
University of Massachusetts Amherst
Amherst, MA 01003
Email: {fabricio,ribeiro,towsley}@cs.umass.edu
† State Key Lab for Manufacturing Systems
Xi'an Jiaotong University
Xi'an P.R.China
Email: phwang@sei.xjtu.edu.cn

Technical Report UM-CS-2012-023

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, network theory, among other areas. 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 hard 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 remains below a 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 class of set size distribution estimation problems. 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 original (pre-sampling) set size distribution 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 size distribution (in packets) of TCP/UDP flows [2], and in next generation Internet capacity planing, such as estimating the number of copies of a movie in a CDN of next-generation routers. Fortunately, simple maximum likelihood [2] 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 nodes without sampled incoming edges are unobservable, 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. The latter result seemly defies intuition. 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 may 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 when it can be avoided. Another interesting property is:

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

The above observation is interesting because power-laws have more tail probability mass and, thus, large sets are more likely to have sampled elements than in 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 dataset, 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 the directed edges are independently sampled with probability p. Henceforth, each person with more than one observed incoming email shall 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 yields *no* reduction in estimation errors.** 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 $N = 50\times10^3$. 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 agreemeent 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 θ_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 5×10^3 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 that we describe in Section IV, the difficulties experienced above arise due to the use of small sampling probability ($p < 0.5$) with heavy-tailed distributions, ant 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 makes the estimates closer to the true in-degree distribution. 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 $\mathcal{S}_k$ be a nonempty set of elements, $k = 1, \ldots, m$, with $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, $i, j = 1, \ldots m$, $i \neq j$. Let $S_k = |\mathcal{S}_k|$ denote the size of the k-th set and assume set sizes are i.i.d. with distribution $S_k \sim \boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$,

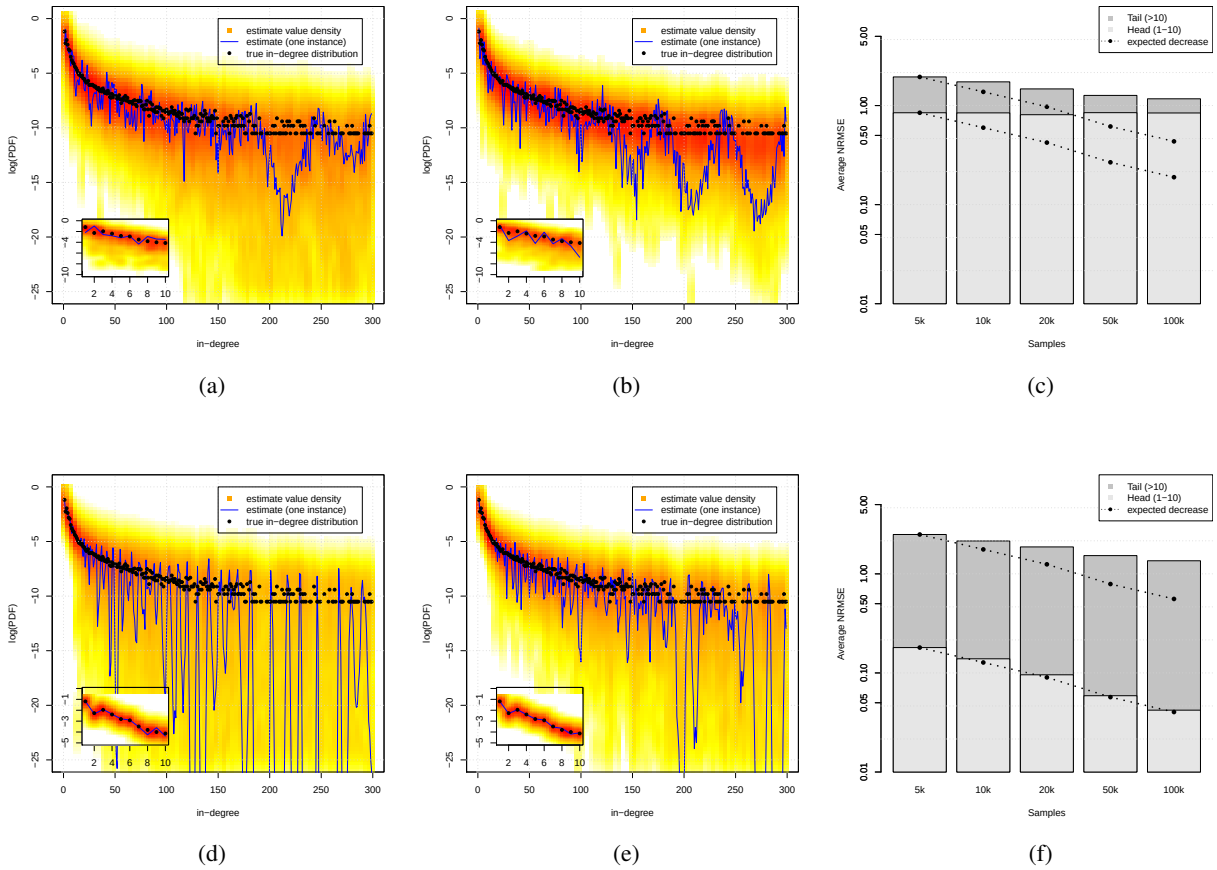


Figure 1. The first row (a-c) shows the results for $p = 0.25$, while the second row (d-e) shows the corresponding plots for $p = 0.90$. (a-b,d-e) True degree distribution, one example of estimate and heat map indicating the ocurrence rates of the estimate values for $N = 10 \times 10^3$ samples (first column) and $N = 50 \times 10^3$ samples (second column), respectively. The red color in the heat map indicates high density of estimated values and yellow (white) indicates low (no) density of estimated values. A subplot shows a zoom-in for the first degrees. (c,f) Average NRMSE of the head and the tail of the distribution for $N \in \{1, 5, 10, 20, 100\} \times 10^3$. Dashed line shows how the error should vary with the number of samples. In (c) we have the **typical behavior of wrong estimates**. Increasing the number of samples does not improve the quality of estimates. On the other hand (f) shows the **typical behavior of correct estimates**. Here increasing the number of samples yields lower estimation errors of the head.

$W > 1$ $k \geq 1$. We assume W finite ($W < \infty$). The model breaks nodes (edges) into groups (sets) and our task in what follows is to characterize those groups from incomplete observation (sample) of these sets. To illustrate the model, consider a directed graph; the set of incoming (outgoing) edges of a node k is represented by $\mathcal{S}_k$, $\boldsymbol{\theta}$ is the indegree (outdegree) distribution, and W is the maximum indegree (outdegree). Another straightforward example is representing IP traffic of a communications network, where k is a TCP flow, $\mathcal{S}_k$ 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 $\mathcal{S}_k$, $k = 1, \ldots, m$, with probability p – a process also known as thinning. Let $\alpha(\mathcal{S}_k)$ be a random function that returns the number of observed elements of $\mathcal{S}_k$. Elements are sampled independently (i.e., the sampling process is Bernoulli) and thus,

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

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., $\mathcal{S}_k$ is said to be observable if $\alpha(\mathcal{S}_k) > 0$. Thus, we denote

$$\mathbb{S} = \{\alpha(\mathcal{S}_k) : \alpha(\mathcal{S}_k) > 0\,,\, k = 1, \ldots, m\}$$

the size of the observable set sizes. Let $N = |\mathbb{S}|$ 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 θ_i is

$$T_i'(\mathcal{S}_1, \ldots, \mathcal{S}_m) = \sum_{k=1}^{m} \frac{\mathbf{1}\{S_k = i\}}{N},$$

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

$$\text{MSE}(T_i'(\mathcal{S}_1, \ldots, \mathcal{S}_m)) = E[(T_i'(\mathcal{S}_1, \ldots, \mathcal{S}_m) - \theta_i)^2] = \frac{\theta_i(1-\theta_i)}{m} \leq \frac{1}{4m}.$$

Thus, for $p = 1$ the estimation error decreases as $1/m$, recalling that m is the number of sets.

Unfortunately, accurately estimating $\boldsymbol{\theta}$ when $p < 1$ is significantly more challenging. Recall that a set $\mathcal{S}_k$ is said to be observable if $\alpha(\mathcal{S}_k) > 0$. *We upfront assume that a unobservable sets cannot be used in the estimation process.* This means that our estimator only has access to sets $\mathcal{S}_k$ where $\alpha(\mathcal{S}_k) > 0$. Here we need another function T_i that takes the observed set sizes $\mathbb{S}$ as inputs and outputs an **unbiased** estimate $T_i(\mathbb{S})$ of θ_i, i.e., $E[T_i(\mathbb{S})] = \theta_i$. In what follows we focus on unbiased estimates; our discussion section (Section VI) extends our results to biased estimators. The Mean Squared Error (MSE) of our estimator is

$$\text{MSE}(T_i(\mathbb{S})) = E[(T_i(\mathbb{S}) - \theta_i)^2].$$

The function T_i that minimizes the MSE with respect to sets of size $i = 1, \ldots, W$ is

$$T_i^\star(\mathbb{S}) = \arg\min_{T_i} \text{MSE}(T_i(\mathbb{S})),$$

s.t. $E[T_i^\star(\mathbb{S})] = \theta_i$.

IV. Results

In this section we present and discuss our results.

Theorem 4.1: Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$ be a distribution where $\exists i_0$ such that $\theta_i \leq 1/2$ for all $i > i_0$. Recall that $N \leq m$ is the number of observed sets out of the total m sets. We show that, as $W \to \infty$, for N sufficiently large any unbiased estimator $T_i(\mathbb{S})$, $i \geq 1$ is such that:

1) When θ_W decreases faster than exponentially in W, i.e., $-\log \theta_W = \omega(W)$, $\text{MSE}(T_i(\mathbb{S})) = O(1/N)$ for $0 < p < 1$.
2) When θ_W decreases exponentially in W, i.e., $\log \theta_W = W \log a + o(W)$ as for some $0 < a < 1$,
   a) $\log[\text{MSE}(T_i(\mathbb{S}))] = \Omega(W/\log N)$, if $p < a/(a+1)$,
   b) $\text{MSE}(T_i(\mathbb{S})) = \Omega(W^{2i+1}/N)$, if $p = a/(a+1)$,
   c) $\text{MSE}(T_i(\mathbb{S})) = O(1/N)$, if $p > a/(a+1)$.
3) When θ_W decreases more slowly than exponential, i.e., $-\log \theta_W = o(W)$,
   a) $\log[\text{MSE}(T_i(\mathbb{S}))] = \Omega(W/\log N)$, if $p < 1/2$,
   b) $\text{MSE}(T_i(\mathbb{S})) = O(1/N)$, if $p \geq 1/2$; more precisely,
      i) $\text{MSE}(T_i(\mathbb{S})) = \omega(1/N)$, if $p = 1/2$ and $\sum_{j=1}^{W} j^{2i}\theta_j = \omega(1)$,
      ii) $\text{MSE}(T_i(\mathbb{S})) = O(1/N)$, if either $p > 1/2$ or $p = 1/2$ and $\sum_{j=1}^{W} j^{2i}\theta_j = O(1)$.

Theorem 4.2: The bounds on the estimation error of the average set size are analogous to the set size distribution bounds.

In what follows we explain how we sketch out the proof of Theorems 4.1 and 4.2 and describe their implications.

A. Lower Bound on Estimation Errors

In this section we derive a lower bound on the Mean Squared Error (MSE) of $T_i(\mathbb{S})$, $i = 1, \ldots, W$. For this we use the Cramér-Rao (CR) lower bound of $T_i(\mathbb{S})$, 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 $\mathcal{S}$ is observed and has j elements is defined as

$$b_{ji}(p) \equiv P\left[\alpha(\mathcal{S}) = j \,|\, \alpha(\mathcal{S}) > 0, |\mathcal{S}| = i\right] = \frac{\binom{i}{j} p^j q^{i-j}}{1 - q^i}, \quad \text{if } 0 < j \leq i \leq W, \tag{1}$$

and $b_{ji}(p) = 0$ otherwise, where $q = 1 - p$. Let $d_j(\boldsymbol{\theta}, p)$ denote the fraction of observed sets with exactly j observed elements. From (1) we have, $j = 1, \ldots, W$,

$$\begin{aligned} d_j(\boldsymbol{\theta}, p) &= P[\alpha(\mathcal{S}) = j |\, |\mathcal{S}| > 0] \\ &= \sum_{i=j}^{W} P[\alpha(\mathcal{S}) = j | \alpha(\mathcal{S}) > 0, |\mathcal{S}| = i] P[|\mathcal{S}| = i |\, \alpha(\mathcal{S}) > 0] \\ &= \sum_{i=j}^{W} b_{ji}(p) \phi_i(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

where

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

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

$$d(\boldsymbol{\theta}, p) = B(p) \boldsymbol{\phi}(\boldsymbol{\theta}),$$

where $d(\boldsymbol{\theta}, p) = (d_1(\boldsymbol{\theta}, p), \ldots, d_W(\boldsymbol{\theta}, p))^{\mathsf{T}}$ and $B(p) = [b_{ji}(p)], j, i = 1, \ldots, W$. To illustrate the distribution $d(\boldsymbol{\theta}, p)$ in our model, note that for a random observed set $\mathcal{S}$,

$$\alpha(\mathcal{S}) \sim d(\boldsymbol{\theta}, p),$$

with likelihood function

$$f(j|\boldsymbol{\theta}) \equiv P[\alpha(\mathcal{S}) = j \,|\, \boldsymbol{\theta}] = (B(p)\boldsymbol{\phi}(\boldsymbol{\theta}))_j = d_j(\boldsymbol{\phi}(\boldsymbol{\theta}), p). \tag{4}$$

In what follows for simplicity we denote $d_j(\boldsymbol{\theta}, p)$ as $d_j(\boldsymbol{\theta})$, $j = 1, \ldots, W$.

Recall that we are interested in functions $T_i(\mathbb{S})$ that take as input the observed subset sizes $\mathbb{S}$ and outputs an unbiased estimate $T_i(\mathbb{S})$ of θ_i, $i = 1, \ldots, W$. Moreover, we want these estimates to be accurate, i.e., MSE$(T_i(\mathbb{S}))$ must be low in respect to θ_i. Otherwise, the estimate is of little use to the practitioner for set sizes of interest, as illustrated in Figure 1.

Thus, it is important to find attainable lower bounds of MSE$(T_i(\mathbb{S}))$. 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 [9, Chapter 2], i.e.,

$$\text{MSE}(T_i(\mathbb{S})) \equiv E[(T_i(\mathbb{S}) - \theta_i)^2] \geq \frac{\left((J^{(\boldsymbol{\theta})}(p))^{-1}\right)_{ii}}{N}, 1 \leq i \leq W. \tag{5}$$

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

$$(J^{(\boldsymbol{\theta})}(p))_{i,k} \equiv \sum_{j=1}^{W} \frac{\partial \ln f(j \,|\, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(j \,|\, \boldsymbol{\theta})}{\partial \theta_k} d_j(\boldsymbol{\phi}(\boldsymbol{\theta})) = \sum_{j=1}^{W} \frac{\partial d_j(\boldsymbol{\phi}(\boldsymbol{\theta}))}{\partial \theta_i} \frac{\partial d_j(\boldsymbol{\phi}(\boldsymbol{\theta}))}{\partial \theta_k} \frac{1}{d_j(\boldsymbol{\phi}(\boldsymbol{\theta}))}, \tag{6}$$

given $\sum_{i=1}^{W} \theta_i = 1$.

The lower bound in (5) is known in the literature as the Cramér-Rao lower bound or *CRLB* for short. Let $T_i^*(\mathbb{S})$ be an unbiased estimator, $i = 1, \ldots$. We say $T_i^*(\mathbb{S})$ is asymptotically efficient if MSE$(T_i^*(\mathbb{S}))$ approaches the Cramér-Rao lower bound in (5) as $N \to \infty$. We show in Appendix D that the Maximum Likelihood Estimator is asymptotically efficient on the set size estimation. The implication of having an efficient estimator is that the lower bounds provided in this paper are tight for N sufficiently large. In what follows we represent $J^{(\boldsymbol{\theta})}(p)$ as $J^{(\boldsymbol{\theta})}$ for simplicity.

B. Obtaining the CRLB

In what follows we derive closed-form lower bounds for the MSE of any unbiased estimator T, as a function of the original set size distribution $\boldsymbol{\theta}$, 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 $J^{(\boldsymbol{\theta})}$ is no easy task as matrix $J^{(\boldsymbol{\theta})}$ is a function of $\partial f(j|\boldsymbol{\theta})/\partial\theta_i$, $i = 1, \ldots, W$, which makes $J^{(\boldsymbol{\theta})}$ a non-linear function of $\boldsymbol{\theta}$. However, observe that the likelihood function $f^\star(j|\boldsymbol{\phi}) \equiv P[\alpha(\mathcal{S}) = j|\boldsymbol{\phi}]$ (where $\mathcal{S}$ is a random observed set) is linear with respect to $\boldsymbol{\phi}$

$$f^\star(j|\boldsymbol{\phi}) \equiv (B\boldsymbol{\phi})_j = d_j(\boldsymbol{\phi}). \tag{7}$$

It is worth noting that $f^\star(j|\boldsymbol{\phi}(\boldsymbol{\theta})) = f(j|\boldsymbol{\theta})$. The Fisher information matrix with respect to $\boldsymbol{\phi}$ is defined as $J^{(\boldsymbol{\phi})} = [J_{i,k}^{(\boldsymbol{\phi})}], i, k = 1, \ldots, W$, where

$$J_{i,k}^{(\boldsymbol{\phi})} \equiv \sum_{j=1}^{W} \frac{\partial d_j(\boldsymbol{\phi})}{\partial \phi_i} \frac{\partial d_j(\boldsymbol{\phi})}{\partial \phi_k} \frac{1}{d_j(\boldsymbol{\phi})}, \tag{8}$$

given $\sum_{i=1}^{W} \phi_i = 1$; and because $d_j(\boldsymbol{\phi})$ is linear in $\boldsymbol{\phi}$, combining (7) and (8) yields

$$(J^{(\boldsymbol{\phi})})^{-1} = B(p)^{-1}\text{diag}(B(p)\boldsymbol{\phi})^{-1}(B(p)^{-1})^\mathsf{T} - \boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}. \tag{9}$$

Here the term $\boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}$ corresponds to the accuracy gain obtained by considering the constraint $\sum_{i=1}^{W} \phi_i = 1$ (see Tune and Darryl [8] for more details and Gorman and Hero [3] for the general formula on adding equality constraints to the CRLB). Quantitatively we can safely ignore the constant term $\boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}$ as we are interested in the behavior of $(J^{(\boldsymbol{\phi})})^{-1}$ as a function of W and the elements of $\boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}$ are typically small. All that is left to do is to find a relationship between $(J^{(\boldsymbol{\phi})})^{-1}$ and $(J^{(\boldsymbol{\theta})})^{-1}$.

We now obtain $(J^{(\boldsymbol{\theta})})^{-1}$ from $(J^{(\boldsymbol{\phi})})^{-1}$ through a multi-variate extension of the single variable chain rule. As $f^\star(j|\boldsymbol{\phi}(\boldsymbol{\theta})) = f(j|\boldsymbol{\theta})$ the chain rule yields

$$\frac{\partial f(j|\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial f'(\phi_j(\theta_i))}{\partial \theta_i} = \frac{\partial f'(\phi_j)}{\partial \phi_j} \cdot \frac{\partial \phi_j(\boldsymbol{\theta})}{\partial \theta_i}, \quad \forall i, j.$$

Using the Jacobian $\nabla H = [h_{ik}]$, $h_{ik} = \partial\theta_k(\boldsymbol{\phi})/\partial\phi_i$ with $\theta_k(\boldsymbol{\phi})$ as given in (3), we arrive at the equivalent multivariate rule [9, pp. 83] to express $(J^{(\boldsymbol{\theta})})^{-1}$ as

$$(J^{(\boldsymbol{\theta})})^{-1} = \nabla H (J^{(\boldsymbol{\phi})})^{-1} \nabla H^\mathsf{T}. \tag{10}$$

Using (9) – detailed derivation relegated to the Appendices – we find:

$$[(J^{(\boldsymbol{\phi})})^{-1}]_{ij} = \sum_{k=\max(i,j)}^{W} \left(\frac{q}{p}\right)^{2k} \binom{k}{j} \binom{k}{i} (-1)^{-i-j}(q^{-i} - 1)(q^{-j} - 1) d_k(\boldsymbol{\theta}). \tag{11}$$

Substituting (11) into (10) – and through a variety of algebraic manipulations detailed in the Appendices – yields

$$[(J^{(\boldsymbol{\theta})})^{-1}]_{ii} = \frac{1}{\eta^2}\Bigg(\underbrace{\frac{1}{(1-q^i)^2}[(J^{(\boldsymbol{\phi})})^{-1}]_{ii}}_{A_1(i)} + \underbrace{\theta_i^2\sum_{j=1}^{W}\sum_{k=1}^{W}\frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{kj}}{(1-q^k)(1-q^j)}}_{A_2(i)} - \underbrace{2\theta_i\sum_{j=1}^{W}\frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{ij}}{(1-q^j)(1-q^i)}}_{A_3(i)}\Bigg), \tag{12}$$

where $\eta = \sum_{j=1}^{W}\phi_j(\boldsymbol{\theta})/(1-q^j)$. Note that term $A_1(i)$ of (12) is proportional to the CRLB of $\boldsymbol{\phi}$, $[(J^{(\boldsymbol{\phi})})^{-1}]_{ii}$ 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 Appendices, we see that $(A_1(i) + A_2(i) - A_3(i))$ grows as a function of $(1-p)/p$ and W, yielding the relation

$$\text{MSE}(T_i(\mathbb{S})) = \Omega\left(\frac{\sum_{j=1}^{W}\left(\frac{1-p}{p}\right)^j\theta_j}{N}\right), \quad i = 1,\ldots,W, \tag{13}$$

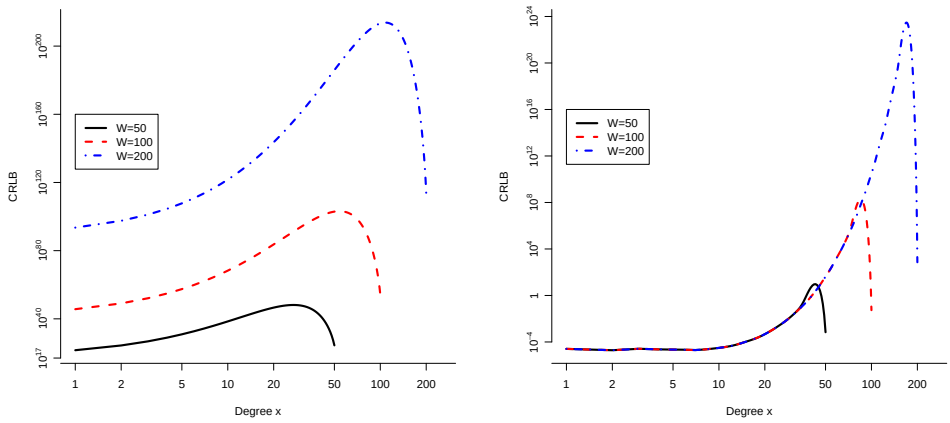
where the number of observed sets N is large but constant in respect to W.

The result in (13) is very powerful as it gives simple estimation error lower bounds as a function of the sampling probability p and the original set size distribution $\boldsymbol{\theta}$. A close look at (13) reveals – a detailed exposition is presented in the Appendices – that when $((1-p)/p)^i\theta_i = \Omega(i^{-1})$ for all $i > i^\star$, $i^\star \ll W$, then the sum in (13) grows at least as fast as the a harmonic series, which grows as $\log W$. On the other hand, we see in the Appendices that when $((1-p)/p)^i\theta_i = O(i^{-\beta})$, $\beta > 1$, then the sum in (13) converges to a constant, more precisely, it grows no faster than a Riemman zeta function with parameter β, $\zeta(\beta)$.

Thus, for a given $\boldsymbol{\theta}$ with $W \gg 1$ the CRLB suffers from an interesting sharp threshold related to the sampling probability p. If p is below this threshold no estimator T_i of θ_i ,$i = 1,\ldots,W$, is able to achieve accurate estimates of θ_i. Below such p threshold, and as long as the number of sampled sets, N, is large enough, there exists estimators $T_i(\mathbb{S})$,$i = 1,\ldots,W$, that can achieve accurate estimates. To be more specific, we look at the threshold behavior of p by breaking down $\boldsymbol{\theta}$ into three broad classes of distributions:

1) If θ_W decreases faster than exponentially in W there is no threshold behavior of p. This is because if $-\log\theta_W = \omega(W)$, then there exists a constant $a < 1$ such that $((1-p)/p)^j\theta_j < a^j$, $j = 1, 2, \ldots$. Hence, the sum in (13) converges to a constant for any $p > 0$, yielding $\text{MSE}(T_i(\mathbb{S})) = \Omega(1/N)$, for $0 < p < 1$. Detailed arguments are presented in the Appendices.
2) If $\log\theta_W = W\log a + o(W)$ then if $p \le a/(a+1)$ yields $((1-p)/p)^j\theta_j = a^{-j}\theta_j = \Omega(1)$, $\forall j$. Hence, the sum in (13) diverges with W. On the other hand, if $p > a/(a+1)$ the sum in (13) converges to a constant. Detailed arguments are presented in the Appendices.
3) Finally, if θ_W decreases more slowly than exponential then if $p = 1/2-\epsilon$, $\epsilon \ge 0$, yields $((1-p)/p)^j > (1+\epsilon/2)^j$, $\forall j$. Hence, because θ_j decreases more slowly than an exponential, the sum in (13) diverges with W. If $p \ge 1/2$ the lower bound in (13) converges to a constant. Detailed arguments are presented in the Appendices.

To illustrate our results, we compute the MSE lower bounds in (12) where $\boldsymbol{\theta}$ is the Enron in-degree distribution 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.



(a) $p = 0.25$ (b) $p = 0.90$

Figure 2. CRLB of the in-degree distribution of the Enron dataset for $N = 10^4$ samples.

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, surprisingly, the accuracy of the average set size follows similar lower bounds of set size distribution estimators T_i, $i = 1, \ldots, W$. We then analyze the accuracy of entropy estimates.

V. Accuracy of Estimated Averages

In this section we focus on the accuracy of the average set size.

A. Average set size

The average set size is $m_{\boldsymbol{\theta}} = \sum_{j=1}^{W} j\theta_j$, or, alternatively, in matrix form

$$m_{\boldsymbol{\theta}} = [1, \ldots, W]\boldsymbol{\theta}^{\mathsf{T}}.$$

Let

$$\frac{\nabla M}{\nabla \theta} = \left[\frac{\partial m_{\boldsymbol{\theta}}}{\partial \theta_1}, \cdots, \frac{\partial m_{\boldsymbol{\theta}}}{\partial \theta_W}\right] = [1, \ldots, W].$$

Let $m(\mathbb{S})$ be an unbiased estimate of the average set size. Using a similar argument used to obtain (10) (see Appendices) yields

$$\begin{aligned} MSE(m(\mathbb{S})) &\geq \frac{\nabla M}{\nabla \theta}(J^{(\boldsymbol{\theta})})^{-1}\frac{\nabla M}{\nabla \theta}^{\mathsf{T}} \\ &= \frac{\nabla M}{\nabla \theta}\left(\frac{\nabla H}{\nabla \phi}(J^{(\boldsymbol{\phi})})^{-1}\frac{\nabla H}{\nabla \phi}^{\mathsf{T}}\right)\frac{\nabla M}{\nabla \theta}^{\mathsf{T}} \\ &= \left(\frac{\nabla M}{\nabla \theta}\frac{\nabla H}{\nabla \phi}\right)(J^{(\boldsymbol{\phi})})^{-1}\left(\frac{\nabla M}{\nabla \theta}\frac{\nabla H}{\nabla \phi}\right)^{\mathsf{T}}. \end{aligned} \tag{14}$$

Note that

$$\begin{aligned}\left[\frac{\nabla M}{\nabla\theta}\frac{\nabla H}{\nabla\phi}\right]_k &= \sum_{i=1}^{W} i h_{ik} \\ &= \sum_{\substack{i=1\\ i\neq k}}^{W} i\left(-\frac{\theta_i}{\eta(1-q^k)}\right) + k\left(\frac{1-\theta_k}{\eta(1-q^k)}\right) \\ &= \frac{1}{\eta(1-q^k)}\left(k - \sum_{i=1}^{W} i\theta_i\right) \\ &= \frac{k-m_\theta}{\eta(1-q^k)}, \end{aligned} \tag{15}$$

where again $\eta = \sum_{j=1}^{W}\phi_j(\boldsymbol{\theta})/(1-q^j)$. Substituting (15) into (14) yields

$$\begin{aligned}\text{MSE}(m(\mathbb{S})) &\geq \frac{1}{N}\sum_{i=1}^{W}\sum_{j=1}^{W}\left(\frac{j-m_\theta}{\eta(1-q^j)}\right)[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}\left(\frac{i-m_\theta}{\eta(1-q^i)}\right) \\ &= \frac{1}{N}\frac{1}{\eta^2}\left(\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{ij[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} + m_\theta^2\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} - \right. \\ &\left. 2m_\theta\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{j[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^i)(1-q^j)}\right) \\ &= \frac{1}{N}\frac{1}{\eta^2}\left(\eta(\sum_{i=1}^{W} i^2\theta_i + \frac{q}{p}m_\theta) + \frac{m_\theta^2}{\theta_j^2}A_2(i) - 2m_\theta\eta(m_\theta + \frac{q}{p}\theta_1)\right), \end{aligned}$$

with $A_2(i)$ as given in (12). Detailed derivations are found in the Appendices. A closer look at $A_2(i)$ reveals

$$A_2(i) = \frac{1}{N}\theta_i^2\left(1+\eta\left(\sum_{j=1}^{W} q^j\theta_j + \sum_{j=1}^{W}\left(\frac{1-p}{p}\right)^j\theta_j\right)\right) = \Omega\left(\frac{\sum_{j=1}^{W}\left(\frac{1-p}{p}\right)^j\theta_j}{N}\right). \tag{16}$$

Note that the lower bound of $m(\mathbb{S})$ in (16) is the same as the lower bound of $T_i(\mathbb{S})$, $i = 1, \ldots, W$, in (13). Hence, a theorem in the lines of Theorem 4.1 can be stated for $m(\mathbb{S})$:

Theorem 5.1: Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$ be a distribution where $\exists i_0$ such that $\theta_i \leq 1/2$ for all $i > i_0$. Recall that $N \leq m$ is the number of observed sets out of the total m sets. We show that, as $W \to \infty$, for N sufficiently large any unbiased estimator of the estimated mean of $\boldsymbol{\theta}$, $m(\mathbb{S})$, must obey the following properties:

1) When θ_W decreases faster than exponentially in W, i.e., $-\log\theta_W = \omega(W)$, $\text{MSE}(m(\mathbb{S})) = O(1/N)$ for $0 < p < 1$.
2) When θ_W decreases exponentially in W, i.e., $\log\theta_W = W\log a + o(W)$ as for some $0 < a < 1$,
   a) $\log[\text{MSE}(m(\mathbb{S}))] = \Omega(W/\log N)$, if $p < a/(a+1)$,
   b) $\text{MSE}(m(\mathbb{S})) = \Omega(W/N)$, if $p = a/(a+1)$,
   c) $\text{MSE}(m(\mathbb{S})) = O(1/N)$, if $p > a/(a+1)$.
3) When θ_W decreases more slowly than exponential, i.e., $-\log\theta_W = o(W)$,
   a) $\log[\text{MSE}(m(\mathbb{S}))] = \Omega(W/\log N)$, if $p < 1/2$,
   b) $\text{MSE}(m(\mathbb{S})) = O(1/N)$, if $p \geq 1/2$; more precisely,
      i) $\text{MSE}(m(\mathbb{S})) = \omega(1/N)$, if $p = 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = \omega(1)$,

ii) $\text{MSE}(m(\mathbb{S})) = O(1/N)$, if either $p > 1/2$ or $p = 1/2$ and $\sum_{j=1}^{W} j^2\theta_j = O(1)$.

Theorem 5.1 states that estimating the average set size is in the same order of hardness as estimating the entire set size distribution.

It is interesting, though, to verify if the same property holds in the case of the average size of the observed sets, i.e., the average set size in respect to $\boldsymbol{\phi}$,

$$m_{\boldsymbol{\phi}} = \sum_{j=1}^{W} j\phi_j.$$

In what follows we show that the difficulty in estimating $m_{\boldsymbol{\phi}}$ is a function of W and is affected only by the first and second moments of $\boldsymbol{\phi}$, that is, as long as $m_{\boldsymbol{\phi}}$ and

$$m_{\boldsymbol{\phi}}^{(2)} = \sum_{j=1}^{W} j^2\phi_j$$

are finite, $m_{\boldsymbol{\phi}}$ can be accurately estimated if enough samples, N, are collected.

Let $\hat{m}_\phi(\mathbb{S})$ denote an unbiased estimate of m_ϕ and let

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

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

$$\begin{aligned}
\text{MSE}(\hat{m}_\phi) &\geq \frac{(1,\ldots,W)(J^{(\phi)})^{-1}(1,\ldots,W)^{\mathrm{T}} - m_\phi^2}{N} \\
&= \sum_{k=1}^{W}\sum_{i=1}^{k}\sum_{j=1}^{k} ij\binom{k}{j}\binom{k}{i}\frac{(-q)^{2k-i-j}}{p^{2k}}(1-q^i)(1-q^j)d_k(\boldsymbol{\phi}) \\
&= \left(\sum_{i=1}^{W}\frac{i(pi + q^{i+1} - 2q^i + q)\phi_i}{p(1-q^i)} - m_\phi^2\right)/N.
\end{aligned}$$

More interestingly, we show that

$$\hat{m}_\phi^\star(\mathbb{S}) = \frac{\sum_{s\in\mathbb{S}} s}{Np} + \left(1 - \frac{1}{p}\right)\frac{\sum_{s\in\mathbb{S}} \mathbf{1}_{s=1}}{N}, \tag{17}$$

is an unbiased efficient (minimum variance) estimator of m_ϕ, yielding

$$\text{MSE}(\hat{m}_\phi^\star(\mathbb{S})) = \left(\sum_{i=1}^{W}\frac{i(pi + q^{i+1} - 2q^i + q)\phi_i}{p(1-q^i)} - m_\phi^2\right)/N.$$

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. Even if we know the sizes of the sampled sets, the mere fact that the set is sampled means that it probably has a higher than average size, as the probability that a set of size i is sampled is $1-(1-p)^i$. Larger variance in the set sizes means larger biases towards sampling larger sets, which in turn makes it harder to unbias these samples.

VI. Discussion

We divide this section in three parts. Section VI-A considers the initialization of estimation procedures. Section VI-B shows that no clever way to process the data $\mathbb{S}$ exists that would allow an estimator to violate the bounds provided in Section IV. Finally, Section VI-C shows that our results can be extended to encompass biased and Bayesian estimators.

A. Initialization of Estimation Procedures

As previously stated, eq. (4) can be used to derive a maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$. 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 to 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 θ_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 draw from the Dirichlet distribution with W parameters $\alpha = (1, \ldots, 1)$, since the Dirichlet PDF at point $\boldsymbol{\theta}$ 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. Therefore, it is possible that EM achieves 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 G 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 θ_i and small variance. For small enough variance, restricting θ_i to be between 0 and 1 does not affect the final estimate significantly and thus the CRLB accurately bounds the MSE.

B. An Application of the Data Processing Inequality

The data processing inequality [10] states that no function of the data may increase the amount of Fisher information already contained in the data. Thus, the bounds in Theorems 4.1 and 5.1 remain unchanged regardless of how the data is pre-processed, no matter how clever the pre-processing approach is. This, of course, encompasses any type of noise filters or machine learning methods.

C. Impact on Different Types of Estimators: Bayesian, Frequentist, 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 $h(\theta_i) = E[T_i(\mathbb{S})] - \theta_i$ be the estimator bias. Then (see for instance Ben-Haim and Eldar [1])

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

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 lower MSE than any unbiased estimator. However, we believe it is unlikely that a large value of $[(J^{(\boldsymbol{\theta})})^{-1}]_{ii}$ (as large as 10^{160} as seen in Section IV-B for the Enron e-mail network) can be compensated by a biased estimator.

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

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

The Fisher information obtained exclusively by the data is $J^{(\boldsymbol{\theta})}$ presented in (6). And the total Fisher information *prior* + *data* is [9, pp. 84]

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

The Cramér-Rao bound of a Bayesian estimator $W_i(\mathbb{S})$ of θ_i with prior $\pi_{\boldsymbol{\theta}}$ yields [9, pp. 85]

$$\text{MSE}(W_i(\mathbb{S})) \geq (J^{(t)})^{-1} = (J^{(p)} + J^{(\boldsymbol{\theta})})^{-1},$$

and thus, if the data contains little Fisher information then a decrease in the MSE is due to the information contained in the prior $\pi_{\boldsymbol{\theta}}$.

VII. Conclusions & Related Work

In this paper we give explicit expressions of MSE lower bounds of unbiased estimators of the distribution of set sizes θ and the average set size $m_{\boldsymbol{\theta}}$ with sampling probability p. We show that the estimation error of $\boldsymbol{\theta}$ 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 there unbiased estimators of some distributions $\boldsymbol{\theta}$ are too inaccurate to be useful for practitioners. Moreover we show that unbiased estimates of m_θ suffer from similar problems.

Not much prior work exists in the literature. Hohn and Veitch [4] 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 one or two estimators. Ribeiro et al. [7] 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 [8] 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, which we rely on in this work.

VIII. Acknowledgments

This research was sponsored by the NSF under CNS-1065133, ARO under MURI W911NF-08-1-0233, and the U.S. Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARO, ARL , or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] Z. Ben-Haim and Y.C. Eldar. On the constrained cramér-rao bound with a singular fisher information matrix. *Signal Processing Letters, IEEE*, 16(6):453 –456, Jun 2009.

[2] Nick Duffield, Carsten Lund, and Mikkel Thorup. Estimating flow distributions from sampled flow statistics. *IEEE/ACM Transactions on Networking*, 13(5):933–946, 2005.

[3] John D. Gorman and Alfred O. Hero. Lower bounds for parametric estimation with constraints. *IEEE Transactions on Information Theory*, 36(6):1285–1301, Nov 1990.

[4] Nicolas Hohn and Darryl Veitch. Inverting sampled traffic. In *IEEE Transactions on Networking*, 2006.
[5] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.
[6] Peter Olver. *Applications of Lie groups to differential equations*. Springer-Verlag, 2nd ed. edition, 2000.
[7] Bruno Ribeiro, Don Towsley, Tao Ye, and Jean Bolot. Fisher information of sampled packets: an application to flow size estimation. In *Proc. of the IMC*, pages 15–26, 2006.
[8] Paul Tune and Darryl Veitch. Fisher information in flow size distribution estimation. In *IEEE Transactions on Information Theory*, volume 57, pages 7011–7035, 2011.
[9] Hary L. van Trees. *Detection, Estimation and Modulation Theory, Part 1*. Wiley, New York, 2001.
[10] Ram Zamir. A Proof of the Fisher Information Inequality via a Data Processing Argument. *IEEE Transactions on Information Theory*, 44(3):1246–1250, 1998.

APPENDIX A
SET SIZE DISTRIBUTION PROOFS

Let $B(p) = [b_{ji}(p)], j, i = 1, \ldots, W$ be a matrix whose elements are given by

$$b_{ji}(p) \equiv P\left[\alpha(\mathcal{S}) = j \,|\, \alpha(\mathcal{S}) > 0, |\mathcal{S}| = i\right] = \frac{\binom{i}{j} p^j q^{i-j}}{1 - q^i}, \quad \text{if } 0 < j \leq i, \tag{18}$$

and $b_{ij}(p) = 0$ otherwise, where $q = 1 - p$.

Lemma A.1 shows a closed formula for the inverse of $B(p)$.

Lemma A.1: $B(p)^{-1} = [b^\star_{ji}(p)]$ $(i, j = 1, \ldots, W)$, where

$$b^\star_{ji}(p) = \begin{cases} \binom{i}{j} p^{-i} (-q)^{i-j} (1 - q^j) & i \geq j \\ 0 & i < j. \end{cases}$$

Proof. Let $B(p)^{-1} = [b^\star_{ji}(p)]$ with $b^\star_{ji}(p)$ defined above. We first show that $Y = B(p)B(p)^{-1}$ is an identity matrix. Consider element (j, i) of Y:

$$y_{ji} = \sum_{l=1}^{W} b_{jl}(p) b^\star_{li}(p)\,. \tag{19}$$

We have three cases: $j > i$, $j = i$, and $j < i$.
Case 1, $j > i$: eq. (19) yields $y_{ji} = 0$ since $b_{jl}(p) = 0$, $\forall l \leq i$ and $b^\star_{li}(p) = 0$, $\forall l > i$.
Case 2, $j = i$: Here $b_{jl}(p) b^\star_{lj}(p) = 0$, $\forall l \neq j$ and (19) yields

$$y_{jj} = \frac{p^j}{1 - q^j} \cdot p^{-j}(1 - q^j) = 1\,.$$

Case 3, $j < i$: eq. (19) yields

$$\begin{aligned} y_{ji} &= \sum_{l=j}^{i} (-1)^{i-l} p^{j-i} q^{i-j} \binom{l}{j} \binom{i}{l} \\ &= p^{j-i} q^{i-j} \sum_{l=j}^{i} (-1)^{i-l} \binom{i}{j} \binom{i-j}{l-j} \\ &= p^{j-i} q^{i-j} \binom{i}{j} \sum_{l=j}^{i} (-1)^{i-l} \binom{i-j}{l-j} \\ &= p^{j-i} q^{i-j} \binom{i}{j} (1-1)^{i-j} \\ &= 0 \end{aligned}$$

Thus, $y_{jj} = 1$, $\forall j$ and $y_{ji} = 0$, $\forall j \neq i$, which concludes our proof. □

Lemma A.1 directly yields the inverse of the Fisher information matrix $J^{(\phi)}$ of a single observed set, as seen in the following lemma.

Lemma A.2: $(J^{(\phi)})^{-1} = [[(J^{(\phi)})^{-1}]_{ij}]$ $(i, j = 1, 2, \ldots, W)$, where

$$[(J^{(\phi)})^{-1}]_{ij} = \sum_{k=\max(i,j)}^{W} \left(\frac{q}{p}\right)^{2k} \binom{k}{j}\binom{k}{i}(-1)^{-i-j}(q^{-i}-1)(q^{-j}-1)d_k(\boldsymbol{\theta}) \tag{20}$$

Proof. Denote $R^{(\boldsymbol{\phi})}(p) = [R^{(\boldsymbol{\phi})}_{ji}(p)] = B^{-1}(p)\text{diag}(B(p)\boldsymbol{\phi})^{-1}$, where $R^{(\boldsymbol{\phi})}_{ji}(p) = b^{\star}_{ji}(p)d_i(\boldsymbol{\phi})$. Based on Lemma A.1 and eq. (2), we have

$$R^{(\boldsymbol{\phi})}_{ji}(p) = \begin{cases} \binom{i}{j}p^{-i}(-q)^{i-j}(1-q^j)d_i(\boldsymbol{\phi}), & i \geq j, \\ 0, & i < j. \end{cases} \tag{21}$$

Since $J^{(\phi)} = R^{(\boldsymbol{\phi})}(p)(B(p)^{-1})^{\mathsf{T}}$, $[(J^{(\phi)})^{-1}]_{ji}$ is computed as the following equation based on Lemma A.1 and eq. (21)

$$\begin{aligned} [(J^{(\phi)})^{-1}]_{ji} &= \sum_{k=1}^{W} R^{(\boldsymbol{\phi})}_{jk}(p)b^{\star}_{ik}(p) \\ &= \sum_{k=\max(i,j)}^{W} \frac{\binom{k}{j}\binom{k}{i}(-q)^{2k-i-j}(1-q^i)(1-q^j)d_k(\boldsymbol{\phi})}{p^{2k}} \\ &= \sum_{k=\max(i,j)}^{W} \left(\frac{q}{p}\right)^{2k} \binom{k}{j}\binom{k}{i}(-1)^{-i-j}(q^{-i}-1)(q^{-j}-1)d_k(\boldsymbol{\phi}) \end{aligned}$$

□

Lemma A.3: $(J^{(\theta)})^{-1} = [[(J^{(\theta)})^{-1}]_{ij}]$ $(i, j = 1, 2, \ldots, W)$, where

$$[(J^{(\theta)})^{-1}]_{ii} = \frac{1}{\eta^2}\left(\frac{[(J^{(\phi)})^{-1}]_{ii}}{(1-q^i)^2} + \theta_i^2 \sum_{j=1}^{W}\sum_{k=1}^{W} \frac{[(J^{(\phi)})^{-1}]_{kj}}{(1-q^k)(1-q^j)} - 2\theta_i \sum_{j=1}^{W} \frac{[(J^{(\phi)})^{-1}]_{ij}}{(1-q^i)(1-q^j)}\right) \tag{22}$$

where $\eta = \sum_{i=1}^{W} \phi_i/(1-q^i)$.

Proof. The relationship between $(J^{(\theta)})^{-1}$ and $(J^{(\boldsymbol{\phi})})^{-1}$ is given by

$$(J^{(\theta)})^{-1} = \nabla H (J^{(\boldsymbol{\phi})})^{-1} \nabla H^{\mathsf{T}}, \tag{23}$$

where $\nabla H = [h_{ik}]$ with $h_{ik} = \partial\theta_k(\boldsymbol{\phi})/\partial\phi_i$. Hence

$$h_{ik} = \begin{cases} -\frac{\phi_i/(\eta(1-q^i))}{\eta(1-q^k)} & i \neq k \\ \frac{1-\phi_i/(\eta(1-q^i))}{\eta(1-q^i)} & i = k \end{cases}$$

where $\eta = \sum_{k=1}^{W} \phi_k/(1-q^k)$ is a constant. Note that from eq. (3) we have $\theta_i = \phi_i/(\eta(1-q^i))$. Therefore

the diagonal elements of $(J^{(\theta)})^{-1}$ can be written as

$$\begin{aligned}
[(J^{(\theta)})^{-1}]_{ii} &= \sum_{j=1}^{W}\sum_{k=1}^{W} h_{ik}[(J^{(\phi)})^{-1}]_{kj}h_{ij}^{\mathsf{T}} \\
&= \sum_{\substack{j=1\\ j\neq i}}^{W}\sum_{\substack{k=1\\ k\neq i}}^{W}\left(-\frac{\theta_i}{\eta(1-q^k)}\right)[(J^{(\phi)})^{-1}]_{kj}\left(-\frac{\theta_i}{\eta(1-q^j)}\right) + \\
&\quad \sum_{\substack{j=1\\ j\neq i}}^{W}\left(\frac{1-\theta_i}{\eta(1-q^i)}\right)[(J^{(\phi)})^{-1}]_{ij}\left(-\frac{\theta_i}{\eta(1-q^j)}\right) + \\
&\quad \sum_{\substack{k=1\\ k\neq i}}^{W}\left(-\frac{\theta_i}{\eta(1-q^k)}\right)[(J^{(\phi)})^{-1}]_{ki}\left(\frac{1-\theta_i}{\eta(1-q^i)}\right) + \left(\frac{1-\theta_i}{\eta(1-q^i)}\right)^2[(J^{(\phi)})^{-1}]_{ii} \\
&= \frac{1}{\eta^2}\left(\frac{[(J^{(\phi)})^{-1}]_{ii}}{(1-q^i)^2} + \theta_i^2\sum_{j=1}^{W}\sum_{k=1}^{W}\frac{[(J^{(\phi)})^{-1}]_{kj}}{(1-q^k)(1-q^j)} - 2\theta_i\sum_{j=1}^{W}\frac{[(J^{(\phi)})^{-1}]_{ij}}{(1-q^i)(1-q^j)}\right). \quad (24)
\end{aligned}$$

□

We split eq. (22) in three parts to carry out its analysis:

$$[(J^{(\theta)})^{-1}]_{ii} = \frac{1}{\eta^2}\left(\underbrace{\frac{[(J^{(\theta)})^{-1}]_{ii}}{(1-q^i)^2}}_{A_1(i)} + \underbrace{\theta_i^2\sum_{j=1}^{W}\sum_{k=1}^{W}\frac{[(J^{(\theta)})^{-1}]_{kj}}{(1-q^k)(1-q^j)}}_{A_2(j)} - \underbrace{2\theta_i\sum_{j=1}^{W}\frac{[(J^{(\theta)})^{-1}]_{ij}}{(1-q^i)(1-q^j)}}_{A_3(i)}\right). \quad (25)$$

A. Analysis of $A_1(i)$

Based on Lemma A.2 and eq. (2), we have

Lemma A.4:

$$A_1(i) = \eta q^{-2i}\sum_{j=0}^{W-i}\binom{i+j}{i}q^{j+i}\theta_{j+i}g_{ij}. \quad (26)$$

where $\eta = \sum_{k=1}^{W}\phi_k/(1-q^k)$ and $g_{ij} = \sum_{k=0}^{j}\binom{i+k}{i}\binom{j}{k}(q/p)^{k+i}$.

Proof.

$$\begin{aligned}
[(J^{(\phi)})^{-1}]_{ii} &= \sum_{k=i}^{W}\left(\frac{q}{p}\right)^{2k}\binom{k}{i}^2(-1)^{-2i}(q^{-i}-1)^2 d_k(\boldsymbol{\phi}) \\
&= \sum_{k=i}^{W}\sum_{j=k}^{W}\left(\frac{q}{p}\right)^{2k}\binom{k}{i}^2(-1)^{-2i}(q^{-i}-1)^2\frac{\binom{j}{k}p^k q^{j-k}\phi_j}{1-q^j} \\
&= (q^{-i}-1)^2\sum_{j=i}^{W}\binom{j}{i}\frac{q^j\phi_j}{1-q^j}\sum_{k=i}^{j}\binom{k}{i}\binom{j-i}{k-i}(q/p)^k \\
&= (q^{-i}-1)^2\sum_{j=0}^{W-i}\binom{i+j}{i}\frac{q^{i+j}\phi_{i+j}g_{ij}}{1-q^{i+j}} \quad (27)
\end{aligned}$$

where $g_{ij} = \sum_{k=0}^{j}\binom{i+k}{i}\binom{j}{k}(q/p)^{i+k}$.

Since $\phi_i/(1-q^i) = \theta_i\cdot\eta$, we can eq. (26) as a function of θ:

$$[(J^{(\phi)})^{-1}]_{ii} = \eta \left(q^{-i} - 1\right)^2 \sum_{j=0}^{W-i} \binom{i+j}{i} q^{i+j} \theta_{i+j} g_{ij}.$$

Therefore

$$A_1(i) = \eta q^{-2i} \sum_{j=0}^{W-i} \binom{i+j}{i} q^{i+j} \theta_{i+j} g_{ij}. \tag{28}$$

□

Lemma A.5: We have the following bounds for $A_1(i)$:

$$A_1(i) < C_i \sum_{k=0}^{i} c_{ik} \sum_{j=0}^{\infty} \mathbf{1}\{k \le j\} (i+j)^{2i} \big(\frac{q}{p}\big)^{i+j} \theta_{i+j} \tag{29}$$

and

$$A_1(i) > C_i c_{ii} \sum_{j=i(i-1)}^{W-i} j^{2i} \big(\frac{q}{p}\big)^{i+j} \theta_{i+j} \tag{30}$$

where

$$C_i = \frac{\eta q^{-i}}{(i!)^2}$$

and

$$c_{ik} = \binom{i}{k} q^k \prod_{l=0}^{i-k-1} (i-l), \quad k = 0, \ldots, i; i = 1, \ldots W.$$

Proof. Since the i-th derivative of $(q/p)^{i+k}$ with respect to q/p, is

$$\frac{\mathrm{d}^i (q/p)^{i+k}}{\mathrm{d}(q/p)^i} = \prod_{l=1}^{i} (k+l)(q/p)^k,$$

we have the following equations for g_{ij}

$$\begin{aligned} g_{ij} &= \frac{1}{i!} \big(\frac{q}{p}\big)^i \sum_{k=0}^{j} \prod_{l=1}^{i} (k+l) \binom{j}{k} (q/p)^k \\ &= \frac{1}{i!} \big(\frac{q}{p}\big)^i \sum_{k=0}^{j} \binom{j}{k} \frac{\mathrm{d}^i (q/p)^{i+k}}{\mathrm{d}(q/p)^i} \\ &= \frac{1}{i!} \big(\frac{q}{p}\big)^i \frac{\mathrm{d}^i \Big(\sum_{k=0}^{j} \binom{j}{k} (q/p)^{i+k}\Big)}{\mathrm{d}(q/p)^i} \\ &= \frac{1}{i!} \big(\frac{q}{p}\big)^i \frac{\mathrm{d}^i \Big((q/p)^i (1+q/p)^j\Big)}{\mathrm{d}(q/p)^i}. \end{aligned}$$

Using a general form of the product rule [6, pp. 318] yields

$$g_{ij} = \frac{1}{i!} \big(\frac{q}{p}\big)^i \sum_{k=0}^{\min\{i,j\}} \binom{i}{k} \big(\frac{1}{p}\big)^{j-k} \prod_{l=0}^{k-1} (j-l) \big(\frac{q}{p}\big)^k \prod_{l=0}^{i-k-1} (i-l), \tag{31}$$

where to simplify the expression we define $\prod_{l=0}^{-1} \cdots = 1$.

Substituting (31) back into (28), we obtain the following expression for $A_1(i)$

$$A_1(i) = C_i \sum_{k=0}^{i} c_{ik} \sum_{j=0}^{W-i} \mathbf{1}\{k \leq j\} \prod_{l=1}^{i}(j+l) \prod_{l=0}^{k-1}(j-l)(q/p)^{i+j}\theta_{i+j} \tag{32}$$

where

$$C_i = \frac{\eta q^{-i}}{(i!)^2}$$

and

$$c_{ik} = \binom{i}{k} q^k \prod_{l=0}^{i-k-1}(i-l), \quad k = 0, \ldots, i; i = 1, \ldots, W.$$

We have the following upper bounds for $A_1(i)$,

$$\begin{aligned} A_1(i) &< C_i \sum_{k=0}^{i} c_{ik} \sum_{j=0}^{W-i} \mathbf{1}\{k \leq j\}(i+j)^{2i}\big(\frac{q}{p}\big)^{i+j}\theta_{i+j} &(33)\\ &< C_i \sum_{k=0}^{i} c_{ik} \sum_{j=0}^{\infty} \mathbf{1}\{k \leq j\}(i+j)^{2i}\big(\frac{q}{p}\big)^{i+j}\theta_{i+j}. &(34)\end{aligned}$$

A lower bound is obtained by noting that

$$\begin{aligned} \prod_{l=1}^{i}(j+l) \prod_{l=0}^{k-1}(j-l) &> j^{i-k} \prod_{l=1}^{k}(j+l) \prod_{l=1}^{k}(j-l+1) \\ &= j^{i-k} \prod_{l=1}^{k}(j^2+j+l-l^2). \end{aligned}$$

The latter is greater than or equal to j^{2i} whenever $j > i(i-1)$ yielding

$$A_1(i) > C_i c_{ii} \sum_{j=i(i-1)}^{W-i} j^{2i}\big(\frac{q}{p}\big)^{i+j}\theta_{i+j}. \tag{35}$$

□

B. Analysis of $A_2(i)$

$$\begin{aligned}
\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{[(J^{(\theta)})^{-1}]_{ij}}{(1-q^i)(1-q^j)} &= \sum_{i=1}^{W}\sum_{j=1}^{W}\sum_{k=1}^{W}\frac{\binom{k}{j}\binom{k}{i}\left(\frac{q}{p}\right)^{2k}(-1)^{-j-i}(q^{-j}-1)(q^{-i}-1)d_k(\boldsymbol{\phi})}{(1-q^j)(1-q^i)} \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi})\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{k}{j}\binom{k}{i}(-q)^{-j-i} \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi})\left(\sum_{i=1}^{k}\binom{k}{i}(-q)^{-i}\right)^2 \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi})\left(\left(-\frac{q}{p}\right)^{-k}-1\right)^2 \qquad \text{using (63)} \\
&= \sum_{k=1}^{W}d_k(\boldsymbol{\phi})-2\sum_{k=1}^{W}\left(-\frac{q}{p}\right)^{k}d_k(\boldsymbol{\phi})+\sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi}) \\
&= 1-2\sum_{k=1}^{W}\left(-\frac{q}{p}\right)^{k}d_k(\boldsymbol{\phi})+\sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi}). \qquad (36)
\end{aligned}$$

First, note that

$$\begin{aligned}
\sum_{k=1}^{W}\left(-\frac{q}{p}\right)^{k}d_k(\boldsymbol{\phi}) &= \sum_{k=1}^{W}\left(-\frac{q}{p}\right)^{k}\sum_{j=1}^{W}\binom{j}{k}p^kq^{j-k}\theta_j\eta \\
&= \eta\sum_{j=1}^{W}q^j\theta_j\sum_{k=1}^{j}\binom{j}{k}(-1)^k \\
&= -\eta\sum_{j=1}^{W}q^j\theta_j. \qquad \text{using (65)} \qquad (37)
\end{aligned}$$

Also,

$$\begin{aligned}
\sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}d_k(\boldsymbol{\phi}) &= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k}\sum_{j=1}^{W}\binom{j}{k}p^kq^{j-k}\theta_j\eta \\
&= \eta\sum_{j=1}^{W}q^j\theta_j\sum_{k=1}^{j}\binom{j}{k}\left(\frac{q}{p}\right)^{k} \\
&= \eta\sum_{j=1}^{W}q^j\theta_j\left(\left(\frac{1}{p}\right)^{j}-1\right) \qquad \text{using (64)} \\
&= \eta\left(\sum_{j=1}^{W}\left(\frac{q}{p}\right)^{j}\theta_j-\sum_{j=1}^{W}q^j\theta_j\right). \qquad (38)
\end{aligned}$$

Replacing eqs. (37) and (38) into (36) yields

$$\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{[(J^{(\theta)})^{-1}]_{ij}}{(1-q^i)(1-q^j)} = 1+\eta\left(2\sum_{j=1}^{W}q^j\theta_j+\sum_{j=1}^{W}\left(\frac{q}{p}\right)^j\theta_j-\sum_{j=1}^{W}q^j\theta_j\right) \quad (39)$$

$$= 1+\eta\left(\sum_{j=1}^{W}q^j\theta_j+\sum_{j=1}^{W}\left(\frac{q}{p}\right)^j\theta_j\right). \quad (40)$$

Therefore,

$$A_2(i)=\theta_i^2\left(1+\eta\left(\sum_{j=1}^{W}q^j\theta_j+\sum_{j=1}^{W}\left(\frac{q}{p}\right)^j\theta_j\right)\right). \quad (41)$$

Note that $A_2(i)$ is positive and may diverge or not depending on the summation $\sum_{j=1}^{W}\left(\frac{q}{p}\right)^j\theta_j$.

C. Analysis of $A_3(i)$

Note that

$$\begin{aligned}\sum_{k=1}^{W}\binom{k}{i}\left(-\frac{q}{p}\right)^k d_k(\boldsymbol{\phi}) &= \sum_{k=i}^{W}\binom{k}{i}\left(-\frac{q}{p}\right)^k\sum_{j=1}^{W}\binom{j}{k}p^kq^{j-k}\theta_j\eta\\ &= \eta\sum_{k=i}^{W}(-1)^k\sum_{j=1}^{W}\binom{j}{i}\binom{j-i}{k-i}q^j\theta_j\\ &= \eta\sum_{j=i}^{W}\binom{j}{i}q^j\theta_j\sum_{k=i}^{j}\binom{j-i}{k-i}(-1)^k\\ &= (-1)^i\eta\sum_{j=i}^{W}\binom{j}{i}q^j\theta_j\sum_{k=0}^{j-i}\binom{j-i}{k}(-1)^k\\ &= (-q)^i\eta\theta_i. \qquad \text{using (66)}\end{aligned} \quad (42)$$

We also have

$$\begin{aligned}\sum_{k=1}^{W}\binom{k}{i}\left(\frac{q}{p}\right)^{2k} d_k(\boldsymbol{\phi}) &= \sum_{k=1}^{W}\binom{k}{i}\left(\frac{q}{p}\right)^{2k}\sum_{j=1}^{W}\binom{j}{k}p^kq^{j-k}\theta_j\eta\\ &= \eta\sum_{k=1}^{W}\left(\frac{q}{p}\right)^k\sum_{j=1}^{W}\binom{j}{i}\binom{j-i}{k-i}q^j\theta_j\\ &= \eta\sum_{j=i}^{W}\binom{j}{i}q^j\theta_j\sum_{k=i}^{j}\binom{j-i}{k-i}\left(\frac{q}{p}\right)^k.\end{aligned} \quad (43)$$

From eq. (42) and (43), we have

$$\sum_{j=1}^{W}\frac{[(J^{(\theta)})^{-1}]_{ij}}{(1-q^j)(1-q^i)} = \eta\theta_i-(-q)^{-i}\eta\sum_{j=i}^{W}\binom{j}{i}q^j\theta_j\sum_{k=i}^{j}\binom{j-i}{k-i}\left(\frac{q}{p}\right)^k \quad (44)$$

and hence,

$$A_3(i) = \underbrace{2\eta\theta_i^2}_{A_{3,1}(i)} \underbrace{- 2\theta_i(-q)^{-i}\eta \sum_{j=0}^{W-i} \binom{i+j}{i} q^{i+j}\theta_{i+j} \sum_{k=0}^{j} \binom{j}{k} \left(\frac{q}{p}\right)^{k+i}}_{A_{3,2}(i)}. \tag{45}$$

Since $A_{3,1}(i)$ is always finite, we only need to compare the magnitude of $A_1(i)$ and $A_{3,2}(i)$. Since $\sum_{k=0}^{j} \binom{j}{k} \left(\frac{q}{p}\right)^{k+i} < g_{ij}$, we can bound $|A_{3,2}(i)|$ by

$$|A_{3,2}(i)| \leq 2\theta_i q^{-i}\eta \sum_{j=0}^{W-i} \binom{i+j}{i} q^{i+j}\theta_{i+j} g_{ij}.$$

Therefore

$$A_1(i) - |A_{3,2}(i)| \geq (q^{-2i} - 2\theta_i q^{-i})\eta \sum_{j=0}^{W-i} \binom{i+j}{i} q^{i+j}\theta_{i+j} g_{ij}.$$

The RHS of the previous inequation is positive when

$$\begin{aligned} q^{-2i} &\geq 2\theta_i q^{-i} \\ \theta_i &\leq \frac{1}{2q^i} < \frac{1}{2}. \end{aligned}$$

Recall that we assumed that $\exists i_0$ such that $\theta_i \leq 1/2$ for all $i > i_0$. Thus by examining only $A_1(i)$ and $A_2(i)$ we can determine whether $[(J^{(\theta)})^{-1}]_{ii}$ diverges or not for $i > i_0$.

APPENDIX B
PROOF OF THEOREM 4.1.

The lower bound of MSE$(T_i(\mathbb{S}))$, given by $[(J^{(\theta)})^{-1}]_{ii}$, is described for each of the three possible cases in Theorem 4.1. The corresponding proofs are shown in what follows.

1) When θ_W decreases faster than exponentially in W.

Proof. Suppose that θ_W decreases faster than exponentially in W. More precisely, assume that $-\log\theta_W = \omega(W)$. It follows that $\log(\theta_W/\theta_{W+1}) \to \infty$ as $W \to \infty$. Hence, for any $\epsilon > 0$, there exists a $W_0(\epsilon)$ such that $\log(\theta_W/\theta_{W+1}) > 1/\epsilon$ for $W > W_0(\epsilon)$. This implies $\theta_{W+1}/\theta_W < e^{-1/\epsilon}$ for $W > W_0(\epsilon)$. Given $p > 0$, we can choose ϵ such that $qe^{-1/\epsilon}/p < 1$. We now apply the ratio test for convergence of an infinite sum to each of the $i+1$ sums in the upper bound for $A_1(i)$ given by (29).

$$\frac{(W+i+1)^{2i}(q/p)^{W+i+1}\theta_{W+i+1}}{(W+i)^{2i}(q/p)^{W+i}\theta_{W+i}} < \frac{(W+i+1)^{2i}}{(W+i)^{2i}}\frac{qe^{-1/\epsilon}}{p}$$

for $W > W_0(\epsilon) - i$ and the latter expression becomes less than one as $W \to \infty$. Hence $A_1(i) = O(1)$ for $0 < p < 1$.

A similar argument can be used to show that $A_2(i) = O(1)$. Hence, $[(J^{(\theta)})^{-1}]_{ii} = O(1)$ for $0 < p < 1$. □

2) When θ_W decreases exponentially in W.

Proof. Suppose that θ_W decreases exponentially in W. More precisely, let $\log\theta_W = W\log a + o(W)$ for $0 < a < 1$. Recall that $A_2(i)$ is positive. Therefore, the logarithm of $[(J^{(\theta)})^{-1}]_{ii}$ in (22) can be lower bounded as follows,

$$\log[(J^{(\theta)})^{-1}]_{ii} \geq \log A_1(i). \tag{46}$$

In addition, the logarithm of $A_1(i)$ in (26) can be bounded by

$$\begin{aligned}\log A_1(i) &\geq W\log(q/p) + \log\theta_W + o(W)\\ &= W\log(qa/p) + o(W)\end{aligned}$$

where the latter equality follows from the hypothesis. Now, if $qa/p > 1$, then $\log A_1(i) = \Omega(W)$, which implies $\log[(J^{(\theta)})^{-1}]_{ii} = \Omega(W)$. Note that $qa/p > 1$ iff $p < a/(a+1)$.

When $p = a/(a+1)$, then $qa/p = 1$. Hence the lower bound of $A_1(i)$ given by (30) is $\Omega(W^{2i+1})$. Hence, $[(J^{(\theta)})^{-1}]_{ii} = \Omega(W^{2i+1})$.

Similarly to the proof for the case where θ_W decreases faster than exponentially in W, we can use the ratio test for convergence of an infinite sum to show that for $qa/p < 1$, $A_1(i) = O(1)$. Hence, it follows that $[(J^{(\theta)})^{-1}]_{ii} = O(1)$ for $p > a/(a+1)$.

□

3) When θ_W decreases slower than exponentially in W.

Proof. Suppose that θ_W decreases slower than exponentially in W. More precisely assume that $-\log\theta_W = o(W)$. The logarithm of $A_1(i)$ can be lower bounded as follows,

$$\begin{aligned}\log A_1(i) &\geq W\log(q/p) + \log\theta_W + o(W)\\ &= W\log(q/p) + o(W)\end{aligned}$$

The latter equality follows from the hypothesis. Now, if $q/p > 1$ (i.e., $p < 1/2$), then $\log A_1(i) \geq \Omega(W)$, which implies $\log[(J^{(\theta)})^{-1}]_{ii} = \Omega(W)$.

When $p \geq 1/2$, it follows that $A_2(i) = O(1)$. In particular if $p = 1/2$ and $\sum_{j=1}^{W} j^{2i}\theta_j = \omega(1)$, we can see from eq. (30) that $A_1(i) = \omega(1)$ and in turn, $[(J^{(\phi)})^{-1}]_{ii} = \omega(1)$.

Note that for $p = 1/2$ each of the $i+1$ sums in the upper bound for $A_1(i)$ given by (29) is bounded by the $2i$-th moment of the set size distribution. Hence, if $\sum_{j=1}^{W} j^{2i}\theta_j = O(1)$, then $[(J^{(\theta)})^{-1}]_{ii} = O(1)$.

Finally, when $p > 1/2$, an argument similar to that used in the case where θ_W decreases faster than exponentially yields $[(J^{(\theta)})^{-1}]_{ii} = O(1)$. □

APPENDIX C
SIMPLIFIED BOUNDS

It is worth noting that $A_2(i)$ gives us a lower bound on $[(J^{(\theta)})^{-1}]_{ii}$, as $A_1(i) - A_3(i) > 0$. Furthermore, the convergence of $A_2(i)$ is given by the convergence of the sum $\sum_{j=1}^{W}(q/p)^j\theta_j$. Therefore, we can write

$$[(J^{(\theta)})^{-1}]_{ii} = \Omega\left(\sum_{j=1}^{W}\left(\frac{1-p}{p}\right)^j\theta_j\right). \tag{47}$$

From that, we derive the following results.

1) When θ_W decreases faster than exponentially in W.

By definition, for any $\epsilon > 0$, there exists a $W_0(\epsilon)$ such that $\log(\theta_W/\theta_{W+1}) > 1/\epsilon$. Given $p > 0$, we can choose ϵ such that $qe^{-1/\epsilon}/p < 1$. The ratio test for convergence of an infinite sum reads

$$\frac{(q/p)^{j+1}\theta_{j+1}}{(q/p)^j\theta_j} < \frac{qe^{-1/\epsilon}}{p} \tag{48}$$

Let $a = qe^{-1/\epsilon}/p$. Hence, there exists a j^* such that for all $j > j^\star$, $((1-p)/p)^j\theta_j < a^j$, $j = 1, 2, \ldots$. Therefore, the sum converges to a constant for any $0 < p < 1$, yielding $[(J^{(\theta)})^{-1}]_{ii} = O(1)$.

2) When θ_W decreases exponentially in W.

By definition, there exists $0 < a < 1$ such that $\log\theta_W = W\log a + o(W)$. When $p \leq a/(a+1)$ it follows that $((1-p)/p)^j\theta_j \geq a^{-j}\theta_j = \Omega(1)$. Therefore, $[(J^{(\theta)})^{-1}]_{ii} = O(W)$. A tighter bound can be obtained by taking into account $A_1(i)$, yielding $\log[(J^{(\theta)})^{-1}]_{ii} = O(W)$ for $p < a/(a+1)$ and $[(J^{(\theta)})^{-1}]_{ii} = O(W^{2i+1})$

for $p = a/(a+1)$. On the other hand, for $p > a/(a+1)$, we have $((1-p)/p)^j\theta_j < a^j\theta_j = O(1)$. Hence, $[(J^{(\theta)})^{-1}] = O(1)$.

3) When θ_W decreases slower than exponentially in W.

When $p < 1/2$, it follows that $(1-p)/p = a > 1$. In this case, there exists a $j^\star$ such that for all $j > j^\star$, $((1-p)/p)^j\theta_j = a^j\theta_j = \Omega(1)$. Hence, $[(J^{(\theta)})^{-1}]_{ii} = O(W)$ for $p < 1/2$. Conversely, when $p > 1/2$, $(1-p)/p = a < 1$. Hence, there exists a $j^\star$ such that for all $j > j^\star$, $((1-p)/p)^j\theta_j = a^j\theta_j = O(1)$. Thus, $[(J^{(\theta)})^{-1}]_{ii} = O(1)$ for $p > 1/2$. At last, for $p = 1/2$, the summation is exactly 1, which also implies $[(J^{(\theta)})^{-1}]_{ii} = O(1)$. In the latter case (i.e., $p = 1/2$), a tigher bound is obtained by taking $A_1(i)$ into account, which yields $[(J^{(\theta)})^{-1}]_{ii} = \omega(1)$ if $\sum j = 1^W j^{2i}\theta_j = \omega(1)$ and $[(J^{(\theta)})^{-1}]_{ii} = O(1)$ if $\sum j = 1^W j^{2i}\theta_j = O(1)$.

APPENDIX D
ASYMPTOTIC EFFICIENCY AND ASYMPTOTIC NORMALITY OF THE MLE $T_i^*(\mathbb{S})$

In this section we show that there exists a Maximum Likelihood Estimator (MLE) $T_i^{(\phi)}(\mathbb{S})$ of ϕ_i that is asymptotic efficient (i.e., $\text{MSE}(T_i^*(\mathbb{S})) = [(J^{(\phi)})^{-1}]_{ii}$) and asymptotic normal. Since the Delta Method is an exact approximation for the Normal distribution, it follows that there exists a MLE $T_i^*(\mathbb{S})$ of θ_i that is asymptotic efficient, which can be obtained by applying the Delta Method to $T_i^{(\phi)}(\mathbb{S})$.

Consider the likelihood function in Eq. (7):

$$f(j|\phi) = \sum_{i=1}^{W} b_{ji}\phi_i.$$

From the sum-to-one contraint on the parameters, it follows that $\phi_1 = 1 - \sum_{i=2}^{W}\phi_i$. Thus we can rewrite the previous eq. as

$$f(j|\phi) = b_{j1} + \sum_{i=2}^{W}(b_{ji} - b_{j1})\phi_i. \tag{49}$$

Hence,

$$\frac{\partial}{\partial\phi_k}\log f(j|\phi) = \frac{b_{jk} - b_{j1}}{b_{j1} + \sum_{i=2}^{W}(b_{ji} - b_{j1})\phi_i} \qquad 2 < k < W.$$

From Theom. 5.1 [5, Chapter 5], we prove that there exists a MLE that is asymptotically efficient and asymptotically normal by showing that assumptions (A0)-(A2) and (A)-(D) are satisfied.

Proof. (A0) Follows from (49).

(A1) The support of ϕ_i for $2 \le i \le W$ is $0 < \phi_i < 1$ subject to $\sum_{i=2}^{W}\phi_i \le 1$.

(A2) Observations are assumed to be independent.

(A3) Follows by the assumption that $0 < \phi_i < 1$ for $2 \le i \le W$.

(A) We have

$$\frac{\partial}{\partial\phi_k}f(j|\phi) = b_{jk}, \quad 2 \le k \le W$$

and hence

$$\frac{\partial^3}{\partial\phi_m\partial\phi_l\partial\phi_k}f(j|\phi) = 0, \quad 2 \le k, l, m \le W.$$

(B) The expectation of the first logarithmic derivative of f is

$$\begin{aligned} E_\phi\left[\frac{\partial}{\partial\phi_k}\log f(j|\phi)\right] &= \sum_{j=1}^{W}\frac{b_{jk}-b_{j1}}{b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i}\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right) \\ &= \sum_{j=1}^{W}b_{jk}-\sum_{j=1}^{W}b_{j1} \\ &= 1-b_{11} \\ &= 0. \end{aligned}$$

As for the second derivative, we have

$$\begin{aligned} E\left[\frac{\partial}{\partial\phi_l}\log f(j|\phi)\frac{\partial}{\partial\phi_k}\log f(j|\phi)\right] &= \sum_{j=1}^{W}\frac{(b_{jl}-b_{j1})(b_{jk}-b_{j1})}{\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right)^2}\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right) \\ &= \sum_{j=1}^{W}\frac{(b_{jl}-b_{j1})(b_{jk}-b_{j1})}{b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} E\left[-\frac{\partial^2}{\partial\phi_l\partial\phi_k}\log f(j|\phi)\right] &= \sum_{j=1}^{W}-\left(-\frac{(b_{jk}-b_{j1})(b_{jl}-b_{j1})}{\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right)^2}\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right)\right) \\ &= \sum_{j=1}^{W}\frac{(b_{jl}-b_{j1})(b_{jk}-b_{j1})}{b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i}. \end{aligned}$$

(C) The vectors $\left[\frac{\partial}{\partial\phi_2}\log f(j|\phi),\frac{\partial}{\partial\phi_3}\log f(j|\phi),\ldots,\frac{\partial}{\partial\phi_W}\log f(j|\phi)\right]$ for $1<j<W$ must be linearly independent with probability 1. Note that and $b_{jk}>0 \iff j\le k$ (in particular, $b_{j1}>0 \iff j=1$). It follows that for $j>k\ge 2$

$$\begin{aligned} \frac{\partial}{\partial\phi_k}\log f(j|\phi) &= \frac{b_{jk}-b_{j1}}{b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i} \\ &= 0, \end{aligned}$$

whereas for $j\le k$,

$$\frac{\partial}{\partial\phi_k}\log f(j|\phi)=\frac{b_{jk}}{\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i}>0.$$

Therefore, the $j-1$ leftmost entries in the j-th vector are 0 while the remainder are positive. Hence the vectors are linearly independent.

(D) Consider a constant $\epsilon_j > 0$ such that $f(j|\phi) = b_{j1} + \sum_{i=2}^{W}(b_{ji} - b_{j1})\phi_i \geq \epsilon_j$ for $1 \leq j \leq W$. Thus,

$$\begin{aligned}\left|\frac{\partial^3}{\partial\phi_m\partial\phi_l\partial\phi_k}f(j|\phi)\right| &= \left|\frac{-(b_{jk}-b_{j1})(b_{jl}-b_{j1})\times 2(b_{jm}-b_{j1})\phi_m(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i)}{\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right)^4}\right| \\ &= \left|\frac{2(b_{jk}-b_{j1})(b_{jl}-b_{j1})(b_{jm}-b_{j1})\phi_m}{\left(b_{j1}+\sum_{i=2}^{W}(b_{ji}-b_{j1})\phi_i\right)^3}\right| \\ &\leq \left|\frac{2(b_{jk}-b_{j1})(b_{jl}-b_{j1})(b_{jm}-b_{j1})\phi_m}{\epsilon_j^3}\right|.\end{aligned}$$

Since $M_{klm}(j) = \left|\frac{\partial^3}{\partial\phi_m\partial\phi_l\partial\phi_k}f(j|\phi)\right| < \infty$, then $E_\phi[M_{klm}(j)] < \infty$ for all k, l, m. □

APPENDIX E
AVERAGE SET SIZE PROOFS

Lemma E.1: Let p be the sampling probability and $\hat{m}_\phi$ denote an unbiased estimate of the average size of the observed sets m_ϕ. Then,

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

Proof. The estimation error lower bound of the average set size is [9, pg.83, Proposition 3]

$$\text{MSE}(\hat{m}_\phi) \geq \frac{(1,\ldots,W)(J^{(\phi)})^{-1}(1,\ldots,W)^{\mathrm{T}} - m_\phi^2}{N}. \tag{50}$$

Lemma A.2 yields

$$\begin{aligned}&(1,\ldots,W)(J^{(\phi)})^{-1}(1,\ldots,W)^{\mathrm{T}} \\ =\ & \sum_{k=1}^{W}\sum_{i=1}^{k}\sum_{j=1}^{k} ij\binom{k}{j}\binom{k}{i}\left(\frac{q}{p}\right)^{2k}(-1)^{2k-i-j}(q^{-i}-1)(q^{-j}-1)d_k(\boldsymbol{\phi}) \\ =\ & \sum_{k=1}^{W}(q/p)^{2k}d_k(\boldsymbol{\phi})\left(\sum_{i=1}^{k} i\binom{k}{i}\frac{q^{-i}-1}{(-1)^i}\right)\left(\sum_{j=1}^{k} j\binom{k}{j}\frac{q^{-j}-1}{(-1)^j}\right) \\ =\ & d_1(\boldsymbol{\phi}) + \sum_{k=2}^{W}(q/p)^{2k}d_k(\boldsymbol{\phi})\left(\left(-\frac{1-q}{q}\right)^k\frac{k}{1-q}\right)^2 \\ =\ & \left(1-\frac{1}{p^2}\right)d_1(\boldsymbol{\phi}) + \frac{1}{p^2}\sum_{k=1}^{W} d_k(\boldsymbol{\phi})k^2.\end{aligned} \tag{51}$$

Now (2) yields

$$d_1(\boldsymbol{\phi}) = \sum_{i=1}^{W}\frac{ipq^{i-1}}{1-q^i}\phi_i \tag{52}$$

and

$$\begin{aligned}\sum_{k=1}^{W} d_k(\boldsymbol{\phi})k^2 &= \sum_{k=1}^{W}\sum_{i=k}^{W} \frac{\binom{i}{k}p^k q^{i-k}}{1-q^i}\phi_i k^2 \\ &= \sum_{i=1}^{W}\sum_{k=1}^{i} \frac{\binom{i}{k}p^k q^{i-k}}{1-q^i}\phi_i k^2 \\ &= \sum_{i=1}^{W}\left(\sum_{k=1}^{i}\binom{i}{k}p^k q^{i-k}k^2\right)\frac{\phi_i}{1-q^i}.\end{aligned}$$

Using the relation

$$\sum_{k=1}^{i}\binom{i}{k}x^k y^{i-k}k^2 = \begin{cases} x, & i=1, \\ ix(ix+y)(x+y)^{i-2}, & i\geq 2.\end{cases}$$

yields

$$\sum_{k=1}^{W} d_k(\boldsymbol{\phi})k^2 = \sum_{i=1}^{W}\frac{ip(ip+q)\phi_i}{1-q^i}. \tag{53}$$

Putting together (50), (51), and (53) yields

$$\text{MSE}(\hat{m}_\phi) \geq \left(\sum_{i=1}^{W}\frac{i(pi+q^{i+1}-2q^i+q)\phi_i}{p(1-q^i)} - m_\phi^2\right)/N \tag{54}$$

which concludes the proof. □

Lemma E.2: Using the observed set sizes $\mathbb{S} = \{\mathcal{S}_k\}_{k=1}^N$ the following

$$\hat{m}_\phi = \frac{\sum_{k=1}^N \mathcal{S}_k}{Np} + \left(1-\frac{1}{p}\right)\frac{\sum_{k=1}^N \mathbf{1}_{\mathcal{S}_k=1}}{N}, \tag{55}$$

is an efficient (smallest variance) unbiased estimator of m_ϕ.

Proof. We start by noting that

$$m_\phi = [1,...,W]\phi = [1,...,W]B^{-1}d(\boldsymbol{\phi}). \tag{56}$$

Denote $z = [z_1,\ldots,z_W] = [1,...,W]B^{-1}$. From Lemma A.1, we have

$$\begin{aligned} z_i &= \sum_{j=1}^{W} j b^\star_{ji} \\ &= \sum_{j=1}^{i} j\binom{i}{j}p^{-i}(-q)^{i-j}(1-q^j) \\ &= (-q/p)^i\sum_{j=1}^{i} j\binom{i}{j}\frac{1-q^j}{(-q)^j} \end{aligned} \tag{57}$$

For $i=1$ (57) yields $z_1 = 1$ and for $2\leq i\leq W$,

$$z_i = (-q/p)^i\left(-\frac{1-q}{q}\right)^i\frac{i}{1-q} = \frac{i}{p}.$$

Therefore,

$$z = \frac{[p,2,3,\ldots,W]}{p}.$$

Thus applying the above back into (56) yields

$$m_\phi = \frac{m_d}{p} + \left(1 - \frac{1}{p}\right) d_1(\boldsymbol{\phi}), \tag{58}$$

where $m_d = \sum_{i=1}^{W} i d_i$ is the expectation of average set size of observed subsets. Rewriting (58) using the set sizes $\mathbb{S}$ we get

$$\hat{m}_\phi = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{\mathcal{S}_k}{p} + \left(1 - \frac{1}{p}\right) \mathbf{1}_{\mathcal{S}_k=1} \right).$$

Based on our assumption that $\{S_k\}_{k=1}^{m}$ is an i.i.d. sequence, we have that $\{\mathcal{S}_k\}_{k=1}^{N}$ is also i.i.d. with distribution $d(\boldsymbol{\phi})$. Therefore,

$$E[\hat{m}_\phi] = E\left[\frac{\mathcal{S}_k}{p} + \left(1 - \frac{1}{p}\right) \mathbf{1}_{\mathcal{S}_k=1} \right],$$

and

$$\text{Var}[(\hat{m}_\phi)^2] = \frac{1}{N} \text{Var}\left[\left(\frac{\mathcal{S}_k}{p} + \left(1 - \frac{1}{p}\right) \mathbf{1}_{\mathcal{S}_k=1} \right)^2 \right].$$

Since

$$E[\mathcal{S}_k] = m_d = \sum_{i=1}^{W} i d_i(\boldsymbol{\phi}),$$

and

$$E[\mathbf{1}_{\mathcal{S}_k=1}] = d_1(\boldsymbol{\phi}),$$

we have $E[\hat{m}_\phi] = m_\phi$ from (58), which indicates that $\hat{m}_\phi$ is unbiased. Then

$$E[(\mathcal{S}_k)^2] = \sum_{i=1}^{W} i^2 d_i(\boldsymbol{\phi}),$$

$$E[(\mathbf{1}_{\mathcal{S}_k=1})^2] = d_1(\boldsymbol{\phi}),$$

and

$$E[\mathcal{S}_k \mathbf{1}_{\mathcal{S}_k=1}] = d_1(\boldsymbol{\phi}),$$

yield

$$\text{Var}[(\hat{m}_\phi)^2] = \frac{\left(1 - \frac{1}{p^2}\right) d_1(\boldsymbol{\phi}) + \frac{1}{p^2} \sum_{k=1}^{W} d_k(\boldsymbol{\phi}) k^2 - m_\phi^2}{N}.$$

From (50) and (51) we find that $\hat{m}_\phi$ is an unbiased estimator that achieves the Cramér-Rao lower bound (i.e., it is an efficient estimator). □

Lemma E.3: Let $\hat{m}$ denote an unbiased estimate of the average set size m_θ. Then,

$$\begin{aligned} \text{MSE}(\hat{m}_\theta) \;\geq\; & \frac{1}{\eta^2} \Bigg(\sum_{i=1}^{W} \sum_{j=1}^{W} \frac{ij[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} + m_\theta^2 \sum_{i=1}^{W} \sum_{j=1}^{W} \frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} - \\ & 2m_\theta \sum_{i=1}^{W} \sum_{j=1}^{W} \frac{j[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^i)(1-q^j)} \Bigg). \end{aligned} \tag{59}$$

Proof.

$$\begin{aligned}\text{MSE}(\hat{m}_\theta) &\geq \frac{\nabla M}{\nabla \theta}\left(\frac{\nabla H}{\nabla \phi}(J^{(\boldsymbol{\phi})})^{-1}\frac{\nabla H}{\nabla \phi}^T\right)\frac{\nabla M}{\nabla \theta}^T \\ &= \left(\frac{\nabla M}{\nabla \theta}\frac{\nabla H}{\nabla \phi}\right)(J^{(\boldsymbol{\phi})})^{-1}\left(\frac{\nabla M}{\nabla \theta}\frac{\nabla H}{\nabla \phi}\right)^T. \end{aligned} \tag{60}$$

where $\frac{\nabla M}{\nabla \theta} = (1, \ldots, W)$. Note that

$$\begin{aligned}\left[\frac{\nabla M}{\nabla \theta}\frac{\nabla H}{\nabla \phi}\right]_k &= \sum_{i=1}^{W} i h_{ik} \\ &= \sum_{\substack{i=1 \\ i\neq k}}^{W} i\left(-\frac{\theta_i}{\eta(1-q^k)}\right) + k\left(\frac{1-\theta_k}{\eta(1-q^k)}\right) \\ &= \frac{1}{\eta(1-q^k)}\left(k - \sum_{i=1}^{W} i\theta_i\right) \\ &= \frac{k-m_\theta}{\eta(1-q^k)}. \end{aligned} \tag{61}$$

Substituting eq. (61) in eq. (60), we have

$$\begin{aligned}\text{MSE}(\hat{m}_\theta) &\geq \sum_{i=1}^{W}\sum_{j=1}^{W}\left(\frac{j-m_\theta}{\eta(1-q^j)}\right)[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}\left(\frac{i-m_\theta}{\eta(1-q^i)}\right) \\ &= \frac{1}{\eta^2}\left(\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{ij[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} + m_\theta^2\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)} - \right. \\ &\quad \left. 2m_\theta\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{j[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^i)(1-q^j)}\right). \end{aligned}$$

□

Similarly to what we did for eq. (22), we split eq. (59) into three pieces to analyze its behavior.

$$\begin{aligned}\text{MSE}(\hat{m}_\theta) &\geq \frac{1}{\eta^2}\left(\underbrace{\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{ij[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)}}_{U_1} + \underbrace{m_\theta^2\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^j)(1-q^i)}}_{U_2} - \right. \\ &\quad \left. \underbrace{2m_\theta\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{j[(J^{(\boldsymbol{\phi})})^{-1}]_{ji}}{(1-q^i)(1-q^j)}}_{U_3}\right). \end{aligned}$$

A. Analysis of U_1

$$\begin{aligned}
\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{ij[(J^{(\phi)})^{-1}]_{ji}}{(1-q^j)(1-q^i)} &= \sum_{i=1}^{W}\sum_{j=1}^{W}\sum_{k=1}^{W} ij\binom{k}{i}\binom{k}{j}\left(\frac{q}{p}\right)^{2k}(-q)^{-i-j}d_k(\boldsymbol{\phi}) \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k} d_k(\boldsymbol{\phi})\left(\sum_{i=1}^{k} i\binom{k}{i}(-q)^{-i}\right)^2 \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k} d_k(\boldsymbol{\phi})\left(\left(-\frac{q}{p}\right)^{-k}\frac{k}{p}\right)^2 \qquad \text{using (62)} \\
&= \frac{1}{p^2}\sum_{k=1}^{W} k^2 d_k(\boldsymbol{\phi}) \\
&= \frac{\eta}{p^2}\sum_{i=1}^{W} ip(ip+q)\theta_i \\
&= \eta\Big(\sum_{i=1}^{W} i^2\theta_i + \frac{q}{p}m_\theta\Big).
\end{aligned}$$

Note that U_1 is bounded by the second moment of the distribution θ.

B. Analysis of U_2

Note that $U_2 = \frac{m_\theta^2}{\theta_i^2}A_2(i)$. Therefore, we conclude that U_2 diverges if either θ_W decreases exponentially in W and $p < a/(a+1)$ or θ_W decreases slower than exponentially in W and $p < 1/2$.

C. Analysis of U_3

$$\begin{aligned}
\sum_{i=1}^{W}\sum_{j=1}^{W}\frac{j[(J^{(\phi)})^{-1}]_{ji}}{(1-q^i)(1-q^j)} &= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k} d_k(\boldsymbol{\phi})\sum_{i=1}^{k}\binom{k}{i}(-q)^{-i}\sum_{j=1}^{k} j\binom{k}{j}(-q)^{-j} \\
&= \sum_{k=1}^{W}\left(\frac{q}{p}\right)^{2k} d_k(\boldsymbol{\phi})\left(\left(-\frac{p}{q}\right)^{k}-1\right)\left(\left(-\frac{p}{q}\right)^{k}\frac{k}{p}\right) \qquad \text{using (63,62)} \\
&= \underbrace{\frac{1}{p}\sum_{k=1}^{W} k d_k(\boldsymbol{\phi})}_{\eta p m_\theta} - \underbrace{\frac{1}{p}\sum_{k=1}^{W}\left(-\frac{q}{p}\right)^{k} k d_k(\boldsymbol{\phi})}_{-\eta q\theta_1} \\
&= \eta(m_\theta + \frac{q}{p}\theta_1).
\end{aligned}$$

Thus,

$$U_3 = 2m_\theta\eta(m_\theta + \frac{q}{p}\theta_1).$$

It is interesting to note that, counterintuitively, U_2 goes to infinity for certain values of p and $\boldsymbol{\theta}$ while U_1 and U_3 are always finite, even though the factor $[(J^{(\phi)})^{-1}]_{ji}$ that appears inside the double summation in U_2 is the same factor that appears multiplied by j and ji in U_1 and U_3, respectively.

D. Proof of Theorem 4.2

Note that U_1, U_2 and U_3 are positive quantities and, moreover, $\text{MSE}(\hat{m}_\theta) > 0 \Rightarrow U_1 + U_2 > U_3$. We observe that U_1 diverges if the second moment of θ is infinite, U_2 diverges if $\sum_{j=1}^{W} \left(\frac{q}{p}\right)^j \theta_j \to \infty$ as $W \to \infty$, while U_3 is always finite.

Proof. **1) When θ_W decreases faster than exponentially in W.**

In this case, the second moment of $\boldsymbol{\theta}$ is finite and the sum $\sum_{j=1}^{W} \left(\frac{q}{p}\right)^j \theta_j = O(1)$ for $0 < p < 1$. Therefore, $\text{MSE}(m(\mathbb{S})) = O(1)$ for $0 < p < 1$.

2) When θ_W decreases exponentially in W.

The second moment of $\boldsymbol{\theta}$ is still finite. However, we can show that the sum $\sum_{j=1}^{W} \left(\frac{q}{p}\right)^j \theta_j$ is $\Omega(W)$ for $p \leq a/(a+1)$ and $O(1)$ for $p > a/(a+1)$ by using an argument similar to the one used in Section E of Appendix A. Hence, $\text{MSE}(m(\mathbb{S})) = \Omega(W)$ for $p \leq a/(a+1)$ and $\text{MSE}(m(\mathbb{S})) = O(1)$ for $p > a/(a+1)$.

3) When θ_W decreases more slowly than exponentially in W.

We can show that the sum $\sum_{j=1}^{W} \left(\frac{q}{p}\right)^j \theta_j$ is $\Omega(W)$ for $p < 1/2$ and $O(1)$ for $p \geq 1/2$ by using an argument similar to the one used in Section E of Appendix A. However, the second moment of $\boldsymbol{\theta}$ shows up in U_1 and it can be either finite or infinite. Although it does not affect the bound for $p < 1/2$, in which case we have $\log \text{MSE}(m(\mathbb{S})) = \Omega(W)$, it does change the bound for $p \geq 1/2$. In particular, if $p = 1/2$ and $\sum_{j=1}^{W} j^2 \theta_j = \omega(1)$, then $\text{MSE}(m(\mathbb{S})) = \omega(1)$. On the other hand, if $p = 1/2$ and $\sum_{j=1}^{W} j^2 \theta_j \geq O(1)$, then $\text{MSE}(m(\mathbb{S})) = \Omega(1)$. Finally, if $p > 1/2$, then $\text{MSE}(m(\mathbb{S})) = \Omega(1)$ as well.

□

APPENDIX F
USEFUL IDENTITIES

$$\sum_{j=1}^{k} j \binom{k}{j} (-q)^{-j} = \left(-\frac{q}{p}\right)^{-k} \frac{k}{p} \tag{62}$$

$$\sum_{j=1}^{k} \binom{k}{j} (-q)^{-j} = \left(-\frac{q}{p}\right)^{-k} - 1 \tag{63}$$

$$\sum_{k=1}^{j} \binom{j}{k} \left(\frac{q}{p}\right)^k = \left(\frac{1}{p}\right)^j - 1 \tag{64}$$

$$\sum_{k=1}^{j} \binom{j}{k} (-1)^k = -1 \tag{65}$$

$$\sum_{k=0}^{j} \binom{j}{k} (-1)^k = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{66}$$