Estimating Network Layer Subnet Characteristics via Statistical Sampling

M. Tozal, Kamil Sarac

Estimating Network Layer Subnet Characteristics via Statistical Sampling

M. Engin Tozal and Kamil Sarac

Department of Computer Science
The University of Texas at Dallas, Richardson, TX 75080 USA
`engintozal,ksarac@utdallas.edu`

Abstract. Network layer Internet topology consists of a set of routers connected to each other through subnets. Recently, there has been a significant interest in studying topological characteristics of subnets in addition to routers in the Internet. However, given the size of the Internet, constructing complete subnet level topology maps is neither practical nor economical. A viable solution, then, is to sample subnets in the target domain and estimate their global characteristics. In this study, we propose a sampling framework for subnets; derive proper estimators for various subnet characteristics including total number of subnets, subnet prefix length distribution, mean subnet degree, and IP address utilization; and analyze the theoretical and empirical aspects of these estimators.

Keywords: network, topology, sampling

1   Introduction

At the network layer, Internet topology consists of a set of routers connected via point-to-point or multi-access links or *subnets*. Most of the existing efforts on capturing a *network layer* map of the Internet focus on *router level* maps [15, 10, 12]. These maps are then studied to understand various topological characteristics of the Internet.

Router level maps typically do *not* consider the nature of the subnets (i.e., point-to-point or multi-access) connecting routers and simply use point-to-point links to represent the connections. On the other hand, subnets are also important building blocks of the Internet topology. An alternative graph representation of the Internet at the network layer may depict subnets as the main entities represented as vertices and consider routers as links connecting those subnets/vertices to each other (see Figure 1). Studying features of *subnet level* maps would improve our understanding of the Internet topology. Among many practical uses of *subnet level* Internet maps are estimating the IP address space utilization in the Internet and developing more representative synthetic Internet graphs.

Despite the benefits of understanding topological properties and evolution of the Internet, drawing complete topology maps turns out to be expensive in terms of time, bandwidth, and computational resources [11, 5]. Therefore, a viable alternative for studying Internet topology is to employ statistical sampling and estimate the parameters of the entire population by analyzing a small unbiased sample of the population. This would naturally require to discover individual subnets and define proper estimators to estimate subnet characteristics. In this paper we present a scheme to achieve this for studying subnet level topological characteristics of the Internet.

In general Internet topology sampling suffer from several practical limitations due to the operational characteristics of the Internet [2, 8]. As an example, in our context we want to sample subnets randomly which requires a list of subnets to

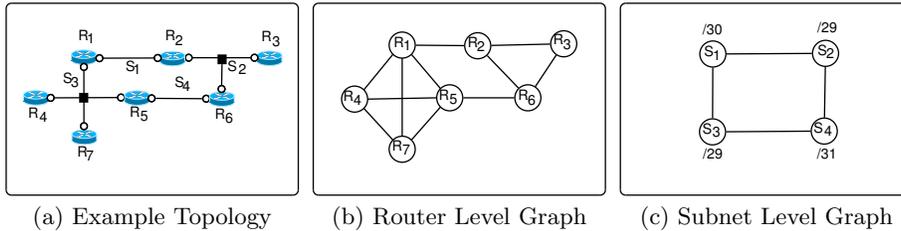| (a) Example Topology | (b) Router Level Graph | (c) Subnet Level Graph |

Fig. 1: A network (left) represented as a router (center) and subnet (right) graph.

be available a priori. Unfortunately, this information is considered confidential by ISPs and therefore, is not available to us, researchers. On the other hand, through active probing we can identify the list of alive IP addresses utilized in these networks. Thus instead of randomly sampling subnets, we can randomly sample IP addresses. In statistical sampling this phenomenon is referred as the discrepancy between selection units (IP addresses) and observation units (subnets) [13]. The problem induced by this discrepancy is that different subnets may utilize different number of IP addresses. As a result random sampling of IP addresses may not indicate random sampling of subnets. In other words, the subnets hosting more IP addresses are more likely to be sampled compared to the ones accommodating less IP addresses. Finally, unresponsive units, i.e., subnets located behind firewalls, or located within rate limiting ISPs, might introduce estimation error depending on whether they are random or not.

Following the above discussion our sampling frame in this study is IP addresses rather than subnets. Therefore, we develop proper estimators for various subnet features using IP addresses of an ISP network as our sampling frame. Specifically, we develop estimators for total number of subnets, subnet mask distribution, mean subnet degree, and IP address space utilization. Our theoretical and empirical evaluations conducted by collecting population characteristics and samples from six geographically disperse Internet service providers demonstrate that our estimators are fairly accurate and stable with small variances. Note that, the scope of this paper is limited to (i) developing an approach for sampling subnets in the Internet; (ii) defining proper estimators for various subnet characteristics; and (iii) demonstrating that these estimators work well in estimating population parameters.

The rest of the paper is organized as follows. Section 2 discusses the challenges of subnet level Internet topology sampling and develop a step-by-step sampling framework for estimating various features of subnets. Section 3 presents our evaluation results on estimator accuracy. Section 4 discusses the related work and Section 5 concludes the paper.

2   Subnet Sampling

In this section we first introduce `exploreNET`, a subnet inference tool that we used in our sampling process. Then, we discuss the challenges of subnet level topology sampling in the Internet and present how we handle these challenges in our sampling framework. Finally, we define a set of subnet characteristics that we are interested in and explain how we can achieve either unbiased or slightly biased estimators with small variance for these characteristics.

2.1   Subnet Inference with ExploreNET

A subnet is a logically visible sub-section of Internet Protocol network (RFC 950) where the connected hosts can directly communicate with each other. At

the network layer, a subnet is independent from its implementation below layer-3 which could be an Ethernet network or an ATM or MPLS based virtual network. For practical purposes we define a subnet S as a set of interfaces (or IP addresses) accommodated by S in addition to a subnet mask (or common IP address prefix length) referring to the IP address range of S.

ExploreNET is a network layer subnet inference tool [16]. Given a target IP address t as input, exploreNET uses active probing to discover the subnet S that contains t along with all other alive IP addresses on S. Furthermore, exploreNET labels S with its observed subnet mask which corresponds to the minimum of the subnet masks encompassing all observed IP addresses of S. To discover the boundaries of a subnet, exploreNET employs a set of heuristics based on hop distance of the subnet being explored, common IP address assignment practices, and routing. Our empirical results show that the proposed subnet inference mechanism achieves 94.9% and 97.3% accuracy rates on Internet2 and GEANT, respectively [14]. Moreover, its success rate is 93% against a ground truth dataset collected by mrinfo in the global public Internet [16]. On the other hand, the probing cost of the tool changes between $2|S|$ and $7|S| + 7$ depending on the configuration of subnet S and its neighboring subnets, where $|S|$ denotes the size of S. Considering that exploreNET is based on a set of heuristics its accuracy rate is subject to change in different network domains and achieving 100% accuracy in every domain is difficult. Since the scope of this paper is limited to statistical subnet sampling, we direct the readers who are interested in the details of subnet discovery to our previous studies [14, 16].

A sampling process has two types of errors, namely, *sampling error* and *nonsampling error* [9]. Sampling errors occur because of chance whereas nonsampling errors can be attributed to many sources including inability to obtain information, errors made in processing data, and respondents providing incorrect information. It is impossible to avoid sampling error, hence, the goal would be to minimize the sampling error by defining unbiased estimators with small variance. On the other hand, nonsampling errors could be eliminated by amending data collection and processing phases. In this context, nonsampling error refers to the error introduced by exploreNET the subnet inference tool that we used in this study and sampling error refers to the error that occurs in our estimations.

Note that our aim in this study is to define good estimators that minimize the sampling error in estimating various subnet features rather than improving subnet inference to eliminate nonsampling error. On the other hand, the sampling approach and estimators proposed in this study do not necessarily depend on exploreNET. Any other future subnet inference tool that collects the subnet accommodating a given IP address should work with our sampling framework.

2.2 Challenges in Subnet Sampling

In this sub-section, we briefly explain the challenges in subnet level Internet topology sampling including the discrepancy between selection and observation units, non-uniform unit sampling, and unresponsive units in sampling.

One challenging issue with sampling is the discrepancy between the selection units and the observation units. Observation units are the main objects that we want to sample and derive their statistical properties, whereas, selection units are the objects in the sampling frame that we can draw from with some probability distribution (preferably uniform). In most cases, selection units and observation units are the same, in some other cases (including Internet topology sampling) however, observation units are not directly available for sampling. To illustrate, we do not have the entire list of the routers or subnets of an ISP to sample from.
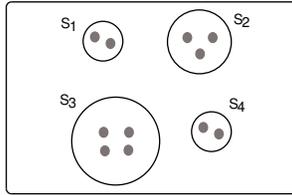
Fig. 2: Subnets hosting more IP addresses are more likely to be sampled compared to the ones with less IP addresses.

The only information available to us is the IP address range of ISPs. As a result, the only sampling frame available to us is the list of alive IP addresses.

A consequence of using IP addresses as sampling frame is non-uniform sampling of subnets. In other words we can uniformly sample from the IP address range of an ISP and give the selected IP addresses to exploreNET as input but the collected subnets would not be sampled uniformly. Remember that exploreNET discovers a subnet as long as one of the IP addresses of the subnet is given as input. As a result subnets with larger degree, i.e., large number of alive IP addresses, appear more frequently in the sampling frame and are observed more frequently compared to the ones having smaller degrees. Put another way, a subnet is drawn with a probability proportional to its degree. Note that we use the term *degree* to denote the number of interfaces of a subnet rather than the number of neighboring subnets of a subnet. Figure 2 shows an example target domain such that large circles depict the subnets and small circles filled in gray denote the IP addresses hosted by the subnets. Considering subnets being sampled through IP addresses, S_3, in Figure 2, is twice as likely to be observed as compared to S_1 because the size of S_1 is half of the size of S_3. As a result, off-the-shelf statistical estimators fail to estimate the characteristics of the subnets due to the unequal drawn probabilities of subnets.

Finally, *unresponsive units* [13] in sampling is another challenge. Most of the tools for collecting topology data in the Internet is based on active probing. However, autonomous systems, especially the ones at the Internet edge, are not completely open to active probing because of security and operational concerns. Some portions of their network are completely behind firewalls or they apply rate limiting in the sense that a router remains silent to probe packets if it is busy or drops probe packets if it suspected a security threat. The interior parts (core) of an ISP network provides a reliable, robust, and efficient communication infrastructure to the other networks. Hence, in this paper we study ISP core networks rather than the edge networks. ISP core networks are more compatible with active probing compared to the edge networks. We also introduce artificial delays and multiple probing in our experiments to minimize the rate limiting factor. Assuming that the unresponsive units in the core network are unmethodical our results are not skewed by them.

2.3 Estimating Subnet Characteristics

In this sub-section we present a generic framework based on Hansen-Hurwitz (HH) Estimator [13] for estimating various subnet features. HH-estimator is an unbiased estimator of population total whenever the selection units are drawn with different probabilities and the sampling is done with replacement. Let $\{y_1, y_2, \ldots, y_j, \ldots, y_n\}$ be a sample of n independent observations from a population of size N. Let p_i be the selection probability of the ith unit in the population such that $\sum_{i=1}^{N} p_i = 1$. HH-estimator estimates the population total

$\tau_y = \sum_{i=1}^{N} y_i$ and it is defined as follows:

$$\widehat{\tau}_y = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{p_j} \tag{1}$$

In Equation (1) dividing observation y_j by its selection probability p_j gives higher weight to the units that are less likely to be selected. An unbiased estimator of population mean is $\widehat{\mu}_y = \widehat{\tau}_y / N$.

In the context of subnet sampling, we define y_i to be the response variable measured on the ith subnet, S_i, of an ISP. Response variables are any characteristic of subnets that we are interested in such as degree, subnet mask, or utilization. Response variables could take discrete values like degree or categorical values like subnet masks. In the former case they are represented as discrete random variables and in the latter case they are represented as indicator random variables. Additionally, the selection probability of a subnet is proportional to its degree, $p_j \propto d_j$, where p_j and d_j are the drawn probability and the degree of subnet S_j, respectively. Then, the drawn probability, p_j, of S_j can be defined in terms of subnet degree as follows:

$$p_j = \frac{d_j}{\sum_{i=1}^{N} d_i} \tag{2}$$

where d_j is the degree of the jth subnet in the sample, d_i is the degree of the ith subnet in the population, and N is the number of subnets in the population. In Equation (2), we neither know the degree, d_i, of each subnet in the population nor the number of subnets, N, in the population in advance. To address this issue, we decided to build a set of alive IP addresses, $F_a = \{a_1, a_2, \ldots, a_k\}$, of the target ISP that we want to estimate its subnet characteristics and utilize the equality between the size of F_a, i.e. $|F_a|$, and the sum of degrees of the observable subnets in the target ISP as $|F_a| = \sum_{i=1}^{N} d_i$. F_a can be obtained as a dataset from LANDER project [5] at USC or be formed in linear probing overhead by pinging each IP address of the target ISP.

Note that, we also use F_a as our sampling frame from which we randomly select n IP addresses with replacement, give each IP address to `exploreNET` as input and obtain the sample set of n subnets. Rearranging the general estimator given in Equation (1) in the context of subnet sampling results in:

$$\widehat{\tau}_y = \frac{|F_a|}{n} \sum_{j=1}^{n} \frac{y_j}{d_j} \tag{3}$$

Equation (3) does not have any unknown term; sampling frame, F_a, is formed prior to sampling process and response variable, y_j, and degree, d_j, are obtained by our subnet inference tool `exploreNET`.

As response variables of subnets, y, are independent from each other, sampling distribution of the point estimator $\widehat{\tau}_y$ is approximately normal by Central Limit Theorem [6], i.e., $\widehat{\tau}_y \sim Normal(\tau_y, \sigma_{\widehat{\tau}_y}^2)$ where

$$\sigma_{\widehat{\tau}_y}^2 = \frac{1}{n|F_a|} \sum_{i=1}^{N} d_i \left(|F_a| \frac{y_i}{d_i} - \tau_y \right)^2 \tag{4}$$

Then we can define a confidence interval (CI) with confidence level $(1-\alpha)100\%$ as $\widehat{\tau}_y \pm z_{\frac{\alpha}{2}} \sqrt{\sigma^2_{\widehat{\tau}_y}}$ such that $z_{\frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the standard normal distribution. Since τ_y is the parameter that we want to estimate and N is usually unknown, an unbiased estimator of the variance of the sampling distribution of $\widehat{\tau}_y$, i.e., $\sigma^2_{\widehat{\tau}_y}$ is defined as follows:

$$s^2_{\widehat{\tau}_y} = \frac{1}{n-1} \left[\frac{|F_a|^2}{n} \sum_{j=1}^{n} \left(\frac{y_j}{d_j} \right)^2 - \widehat{\tau}^2_y \right] \tag{5}$$

Here, we can replace $\sigma^2_{\widehat{\tau}_y}$ with $s^2_{\widehat{\tau}_y}$ in confidence interval construction.

2.4 Important Subnet Characteristics

Although we have derived a general formula for estimating population total over any subnet response variable y in Section 2.3, using Equation (3) requires further arrangements and insights for different subnet features that we want to estimate. In this part we define a set of most relevant subnet characteristics including subnet population size, total subnets with a certain subnet mask, average subnet degree, and IP address utilization percentage and show how to deduce their proper point estimators. Note that this set is not complete and one may suggest new subnet characteristics in the future.

Subnet Population Size (N) Subnet Population Size, N, refers to the number of distinct subnets in a particular ISP. Since $\widehat{\tau}_y$ corresponds to the population total over the response variable y, setting $y_j = 1$ in Equation (3) gives us an unbiased estimator, $\widehat{N}$, of total number of subnets, N, in the population as:

$$\widehat{N} = \frac{|F_a|}{n} \sum_{j=1}^{n} \frac{1}{d_j} \tag{6}$$

Total Subnets with Prefix Length /p (τ_p) Observable prefix length (subnet mask) of a subnet refers to the the length of the initial block of bits common to all IP addresses in the subnet. A subnet having prefix length p is said to be a /p subnet. In order to estimate the total number of subnets of /p let the the response variable measured on subnet S_j, i.e., y_j be an indicator random variable such that

$$y_j = \begin{cases} 1 \ if \ S_j \ is \ a \ /p \ subnet \\ 0 \ otherwise \end{cases} \tag{7}$$

Then we can estimate the total number of subnets having /p prefix, τ_p, in an ISP as follows:

$$\widehat{\tau}_p = \frac{|F_a|}{n} \sum_{j=1}^{n} \frac{y_j}{d_j} \tag{8}$$

Note that, the variance of the estimator τ_p increases as the prefix gets larger (subnet degree gets smaller). The implication of this raise is wider confidence intervals for larger prefix lengths.

In case we want to estimate the prefix length distribution of the subnets we can resort to the fact that $r_p = \tau_p/N$ where r_p is the ratio of subnets having prefix length $/p$. If N is known then $\widehat{r}_p = \widehat{\tau}_p/N$ would be an unbiased estimator of r_p. On the other hand, if N is unknown $\widehat{r}_p = \widehat{\tau}_p/\widehat{N}$ would be a biased estimator of r_p. Nevertheless our experimental results show that the bias in the latter case is small. One should note that, a slightly biased estimator having a small variance of a population parameter is usually preferable compared to an unbiased estimator having a large variance.

Mean Subnet Degree ($\boldsymbol{\mu_d}$) Degree of subnet S_j, d_j, corresponds to the number IP addresses accommodated by S_j. To calculate the average degree throughout an entire ISP we need to compute the total of all degrees divided by the total number of subnets. An estimator of mean degree, μ_d, is defined as $\widehat{\mu}_d = |F_a|/\widehat{N}$

Note that, although $\widehat{N}$ is an unbiased estimator of N, $\widehat{\mu}_d$ is not an unbiased estimator of μ_d because $E[\widehat{\mu}_d] \neq |F_a|/E[\widehat{N}]$. However, our experimental results show that the bias and variance of $\widehat{\mu}_d$ is extremely small and we get very accurate and stable estimations of μ_d.

IP Address Space Utilization ($\mathcal{U}$) `ExploreNET` annotates the subnets with their observed prefix lengths while discovering their in-use IP addresses. Prefix length, p, of subnet S_j indicates the observable IP address capacity, c_j, of S_j while its degree d_j indicates the number of IP addresses that have been utilized. Capacity of subnet S_j is defined as

$$c_j = \begin{cases} 2 & \textit{if } S_j \textit{ is a } /31 \textit{ subnet} \\ 2^{32-p} - 2 & \textit{if } S_j \textit{ is a } /p \textit{ subnet} \\ & \textit{such that } p \leq 30 \end{cases} \qquad (9)$$

ISP utilization $\mathcal{U}$ is defined as the total number of alive IP addresses divided by the total capacity. Again total number of alive IP addresses corresponds to our sampling frame size, $|F_a|$. On the other hand, we estimate total capacity τ_c as $\widehat{\tau}_c = \sum_{\forall p} c_p \widehat{\tau}_p$ where c_p is the capacity of a subnet having prefix length $/p$ and $\widehat{\tau}_p$ is the estimation of the total number of subnets having prefix length $/p$. Since $E[\widehat{\tau}_c] = \tau_c$, $\widehat{\tau}_c$ is an unbiased estimator of total capacity τ_c. Then we can define ISP utilization as $\widehat{\mathcal{U}} = |F_a|/\widehat{\tau}_c$. Again, $\widehat{\mathcal{U}}$ is not an unbiased estimator of $\mathcal{U}$ but our experimental results demonstrate that it is a very good estimator with small bias and variance.

3 Evaluations

In this section, we evaluate sampling errors introduced by our estimators. For this, we need a sampling frame consisting of alive IP addresses clustered in a number of subnets in a network. Such a network could be generated synthetically; could be collected from the Internet using subnet inference techniques; or could be a genuine network available on its web page or obtained from its network operator. Synthetic networks may not be interesting and genuine topologies are available for a few small sized research networks including Internet2 and GEANT which are not large enough for sampling [1]. Therefore, in our evaluations we decided to work with collected topologies. To the best of our knowledge there is no subnet level topology data available for a large sized network. Hence, we use `exploreNET` to conduct a *census* of subnets over six geographically dispersed medium sized ISPs [16] including PCCW Global (ISP-1), nLayer (ISP-2),

France Telecom (ISP-3), Telecom Italia Sparkle (ISP-4), Interroute (ISP-5), and MZIMA (ISP-6). The collected population topology via census might posses non-sampling error introduced by the employed subnet inference tool, i.e., collected topology might slightly differ from the underlying genuine topology. In our sampling we use the same subnet inference tool to conduct our survey of subnets, thus, nonsampling error in the census propagates to the survey. Considering that nonsampling error in census and survey cancel each out while calculating sampling error of point estimates, our evaluations in this section are not affected by nonsampling error. On the other hand, using our methodology for a survey in the Internet will produce results that would contain both sampling and nonsampling errors. The magnitude of those nonsampling errors would depend on the accuracy of the subnet inference tool used in sampling (see Section 2.1).

First, we identified the IP address space for each ISP. Then, using an AS relationship dataset provided by CAIDA, we removed the ranges of IP addresses that are assigned to their customer domains. This pre-processing step enables us to focus on the core of ISP networks excluding the topology information of their customer domains which are managed and operated by others. Next, we utilized active probing to identify alive IP addresses which form our sampling frame, F_a. Table 1 shows the sampling frame sizes, $|F_a|$, of our target ISPs.

Table 1: Sampling Frame Sizes, $|F_a|$, for target ISPs

ISP-1	ISP-2	ISP-3	ISP-4	ISP-5	ISP-6	Σ
45,018	54,636	17,170	8,380	21,209	16,453	162,866

When we collected the entire subnet population for six ISPs having $162,866$ IP addresses in total, `exploreNET` successfully determined a subnet for $155,309$ of them. On the other hand, it failed to explore a proper subnet for $7,557$ of these IP addresses and returned them as /32 subnets. In fact, none of the IP addresses within /29 proximity of any of these $7,557$ IP addresses has responded to the probes sent out by `exploreNET`.

We then randomly selected 10% of the IP addresses from the sampling frame of each ISP and estimated population parameters including population size, N, prefix length distribution, τ_p, mean subnet degree, μ_d, and IP address space utilization, $\mathcal{U}$.

Note that, correcting the affect of singular IP addresses for which `exploreNET` failed to return a subnet requires special care in our estimation process. For example, including /32 samples into the estimation causes underestimated mean subnet degree, $\widehat{\mu}_d$. On the other hand, excluding them causes underestimated population size, $\widehat{N}$. Specifically, we include /32 subnets in population size estimation, $\widehat{N}$, because each one represents a subnet and we exclude them from other estimators because we assume that these /32 subnets distributed randomly, i.e., they could be any size subnet in the underlying topology and their omission does not skew the results in any way. The exclusion of /32 subnets is not simply ignoring them in the estimation process but estimating the number of such subnets and subtracting it from the estimated total number of subnets.

Remember that, the scope of this paper is confined to deriving proper estimators for subnet properties in the Internet and validating them, hence, in the rest of this section, we present the estimations, compare them with population parameters, and interpret the discrepancy between them. We plan to conduct another study based on sampled data to analyze and interpret the motifs and patterns of subnet characteristics in the Internet as a future work.

3.1 Subnet Population Size (N)

Remember that population size refers to total number of subnets appearing in an ISP. Table 2 shows the population size, N, and the corresponding point estimate, $\widehat{N}$, for six ISPs. Third column of the table shows sampling error (SE), $\widehat{N} - N$. The values in the table includes /32 singular IP addresses because these IP addresses belong to some subnet that `exploreNET` could not discover.

Table 2: Population subnet sizes and estimations for all ISPs

	N	$\widehat{N}$	SE	E
ISP-1	4171	4167.53	3.47	247.99
ISP-2	1688	1625.14	62.86	156.38
ISP-3	8809	8791.65	17.35	94.36
ISP-4	3172	3272.01	100.01	116.55
ISP-5	7687	7904.60	217.6	265.44
ISP-6	3119	3180.68	61.68	181.02

Note that the point estimates, $\widehat{N}$, and sampling error, SE, would change for each sample taken from the population. On the other hand, fourth column of the table, E, shows the maximum error of the point estimate for $\widehat{N}$ with 90% confidence level. E is calculated as $z_{0.05}s_{\widehat{N}}$ where $s_{\widehat{N}}$ is the estimated standard deviation of the sampling distribution of $\widehat{N}$ and $z_{0.05}$ is the z-score of the 0.05 right tail probability of standard normal distribution and it demonstrates the fact that 90% of the time the error of the point estimate, $\widehat{N}$, would be less than E. The high deviation of E for some ISPs is because of the variability of the factor $|F_a|/d_j$ in Equation (5) where $y_j = 1$. In other words, HH estimator will have low variance as there is less variability among the values of y_j/p_j with the extreme case y_j/p_j values are exactly proportional in Equation (1) [13].

3.2 Total Subnets with Prefix Length /p (τ_p)

Prefix length (subnet mask) observed by `exploreNET` is the longest prefix length encompassing all alive IP addresses of a given subnet. Consequently, observed prefix length distribution is a natural way of grouping subnet degrees into bins where the range of the bins grow from larger prefix lengths to the smaller. In this part, we evaluate how well the sample prefix length distribution conforms to the population prefix length distribution for six ISPs. Table 3 presents total number of subnets having /p prefix length, τ_p, its estimated value, $\widehat{\tau}_p$, the prefix length probability mass function (p.m.f.), r, and its estimation, $\widehat{r}$, for PCCW Global (ISP-1). In the estimation process we omitted IP addresses that `exploreNET` failed to return a subnet, i.e., /32 subnets. Assuming that these /32 subnets are randomly distributed over the entire population, their exclusion does not affect the accuracy of the estimated prefix length p.m.f, $\widehat{r}$.

In our data collection process, we encountered a small number of very large subnets such as the ones having /20, /21, or /22 prefix lengths with at least one thousand IP addresses. A close examination of randomly selected IP addresses from these subnet ranges via DNS name resolution queries revealed the fact that almost all of these large subnets belong to Akamai Technologies, an online content distribution service provider with a global presence in the Internet. An interesting detail in this case is that according to DNS records, those IP addresses belonged to Akamai Technologies while the IP-to-AS mapping dataset that we used mapped those IP addresses to their hosting ASes. Given that the main focus of this work is on building statistical estimations, we leave the close analysis of this type of interesting cases to our future study where we plan to study the data to understand their implications.

Table 3: PCCW Global (ISP-1) prefix length distribution and its estimation

	τ_p	$\widehat{\tau}_p$	r_p	$\widehat{r}_p$
/20	3	2.94708	0.001266	0.001245
/21	3	2.83139	0.001266	0.001196
/22	7	7.3299	0.002954	0.003096
/23	3	2.62722	0.001266	0.001110
/24	24	24.8725	0.010127	0.010507
/25	25	24.9367	0.010549	0.010534
/26	123	130.435	0.051899	0.055101
/27	152	145.53	0.064135	0.061478
/28	262	284.222	0.110549	0.120066
/29	440	426.242	0.185654	0.180061
/30	899	930.165	0.379325	0.392938
/31	429	385.068	0.181013	0.162668

Although Table 3 demonstrates that the estimated values are close to the population parameters, we need to apply goodness-of-fit test [6] to validate whether the estimated prefix length p.m.f. conforms to the population prefix length p.m.f.. The test statistic for goodness-of-fit test is χ^2 and it is defined as:

$$\chi^2 = \sum_p \frac{(\widehat{\tau}_p - E_p)^2}{E_p}, \qquad \forall p \qquad (10)$$

where p is prefix length, $\widehat{\tau}_p$ is estimated number of subnets having prefix length $/p$, and $E_p = r_p \sum_{p=0}^{31} \widehat{\tau}_p$ is the expected number of subnets having prefix length $/p$ under the population distribution. Null hypothesis, H_0, is the assertion that the estimated and population prefix length distributions are the same, i.e., $H_0 : \{\forall p, \widehat{r}_p = r_p\}$. Whereas, alternative hypothesis, H_1, is the opposite of H_0, i.e., $H_1 : \{\exists p, \widehat{r}_p \neq r_p\}$.

Computing the χ^2 statistic for PCCW Global results in $X^2 = 8.6591$ with degrees of freedom (df) 9 and its related $p\text{-}value$ is 0.4693. Since $p\text{-}value$ is greater than the significance level $\alpha = 0.01$ we conclude that there is not enough evidence to reject the null hypothesis, i.e., estimated prefix length distribution conforms to the population prefix length distribution.

Table 4: $\mathbf{X^2}$ scores and $p\text{-}values$ for all ISPs

	df	$\mathbf{X^2}$	p-value
ISP-1	9	8.659119	0.469317
ISP-2	8	11.53715	0.17308
ISP-3	4	5.513502	0.238545
ISP-4	6	2.1376	0.906618
ISP-5	8	2.02614	0.9802
ISP-6	7	4.97965	0.662447

Instead of tabulating population and estimated prefix length distributions for other five ISPs we present goodness-of-fit test results in Table 4. Since the $p\text{-}values$ of all ISPs in Table 4 are greater than the significance level $\alpha = 0.01$, we again conclude that there is not enough evidence to reject the null hypothesis asserting that the estimated prefix length distribution conforms to the population prefix length distribution for all ISPs. As a result, prefix length ratio estimator, $\widehat{r}_p$, estimates the population prefix length ratio r_p well in all of our case studies.

3.3 Mean Subnet Degree (μ_d)

Remember that mean subnet degree estimator $\widehat{\mu}_d$ is a biased estimator of the population mean subnet degree μ_d. Even tough the first two columns of Table 5 show that the estimated value $\widehat{\mu}_d$ is pretty close to the population value μ_d, we need to make sure that the maximum error of the point estimator, $\widehat{\mu}_d$, is not much and we can get good estimates of the mean subnet degree most of the time.
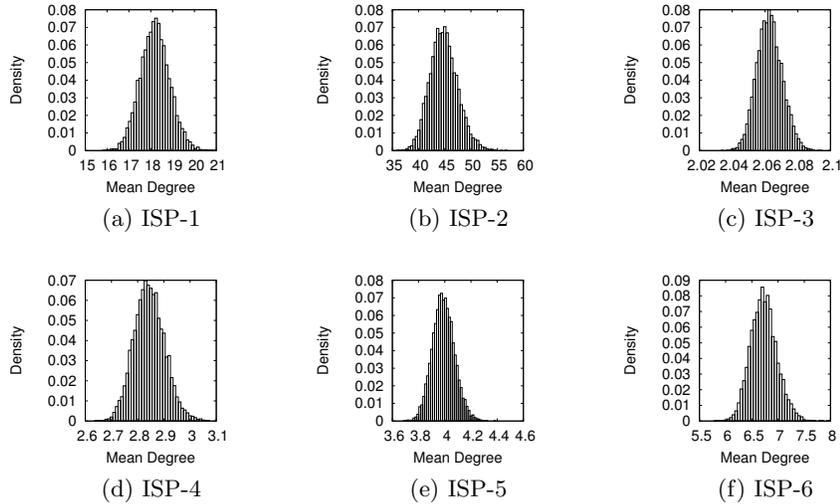
Fig. 3: Mean subnet degree $(\widehat{\mu}_d)$ sampling distribution for all ISPs obtained via 10000 instances of Monte Carlo simulation

Table 5: Mean subnet degree figures for all ISPs

	$\mu_{\mathbf{d}}$	$\widehat{\mu}_{\mathbf{d}}$	**Min**	**Max**	**E**	**NE**
ISP-1	18.235	18.257	15.956	20.975	1.128	1.685
ISP-2	44.904	46.112	36.025	57.12	4.536	1.716
ISP-3	2.063	2.066	2.034	2.095	0.013	1.7
ISP-4	2.849	2.797	2.647	3.079	0.098	1.672
ISP-5	3.997	4.049	3.696	4.39	0.144	1.691
ISP-6	6.742	6.224	5.846	7.986	0.437	1.74

Since variance of the sampling distribution of $\widehat{\mu}_d$ has no closed form solution we resorted to Monte Carlo simulation to determine a confidence interval for μ_d.

We took 10000 independent samples from the population database and estimated mean subnet degree of the sample. Figure 3 shows the histograms demonstrating the approximate sampling distribution of $\widehat{\mu}_d$. The third, *Min*, and fourth, *Max*, columns of Table 5 present the minimum and the maximum of $\widehat{\mu}_d$ point estimations obtained over 10000 samples of the same size (10%), respectively. Although the minimum and maximum estimates fall a little bit off the population mean subnet degree, the bell shape of the distributions suggest that they are outliers. To have a better idea we constructed a confidence interval with confidence level 90% and computed maximum error of the point estimate. Column five, *E* shows the maximum error that could be obtained 90% of the time while estimating the mean subnet degree. Since each ISP has a different mean subnet degree and variance we normalized the maximum error of estimate for $\widehat{\mu}_d$, *E*, by dividing it with the estimated standard deviation of $\widehat{\mu}_d$ as shown in the sixth column, *NE*. The *NE* values suggest that for an ISP, 90% of the time the mean subnet degree estimation conveys a maximum error of approximately 1.7 of standard deviation.

Note that, the ISPs having large mean subnet degree are those ones that home very large network layer subnets belonging to Akamai Technologies.

3.4   IP Address Space Utilization ($\mathcal{U}$)

Similar to mean subnet degree, IP address space utilization estimator, $\widehat{\mathcal{U}}$, is a biased estimator of population IP address space utilization, $\mathcal{U}$. The first and sec-

ond columns of Table 6 presents that the estimated utilization, $\widehat{\mathcal{U}}$, is not far from the population utilization $\mathcal{U}$. However, this is a single instance of sampling and to have a better idea on whether our slightly biased estimator works well most of the time we employed Monte Carlo simulation. Again we repeated sampling for 10000 times from our population database and estimated the value of $\mathcal{U}$ for each sample.
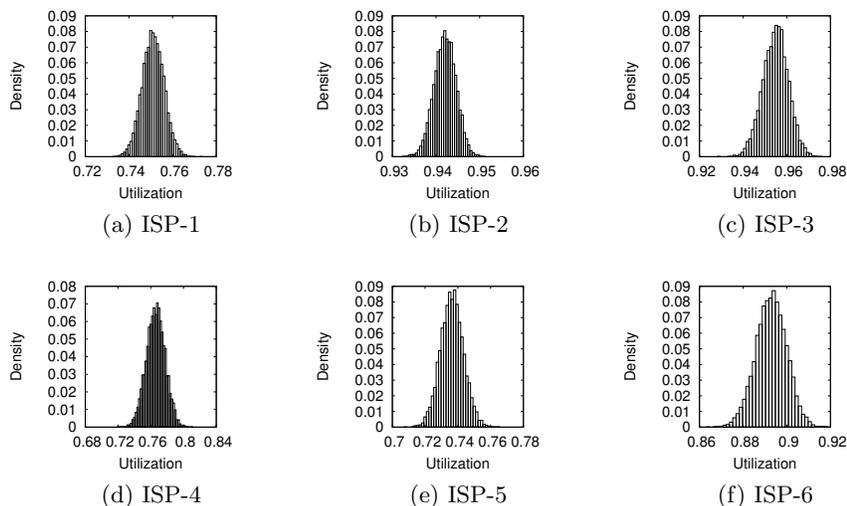

(a) ISP-1   (b) ISP-2   (c) ISP-3

(d) ISP-4   (e) ISP-5   (f) ISP-6

Fig. 4: Utilization ($\widehat{\mathcal{U}}$) sampling distribution for all ISPs obtained via 10000 instances of Monte Carlo simulation

Table 6: Utilization figures for all ISPs

	$\mathcal{U}$	$\widehat{\mathcal{U}}$	Min	Max	E	NE
ISP-1	0.752	0.748	0.734	0.774	0.008	1.645
ISP-2	0.942	0.944	0.933	0.951	0.004	1.654
ISP-3	0.955	0.953	0.929	0.977	0.01	1.691
ISP-4	0.767	0.775	0.722	0.811	0.019	1.653
ISP-5	0.737	0.747	0.709	0.766	0.012	1.656
ISP-6	0.894	0.89	0.865	0.919	0.012	1.67

Figure 4 shows the sampling distribution of $\widehat{\mathcal{U}}$ obtained over 10000 samples of the same size (10%) for six ISPs. Third and fourth columns of Table 6 present the minimum and maximum utilization estimates obtained by Monte Carlo simulation. However, the bell shape of the distributions in Figure 4 suggests that the minimum and maximum values are not likely to occur frequently. To have an idea on the error of the estimator we constructed a confidence interval with 90% confidence level. Column six, E, demonstrates the maximum error of $\widehat{\mathcal{U}}$ at confidence level 90%. To put in other words, it says that 90% of the time the maximum error of sample utilization, $\widehat{\mathcal{U}}$, is E. Again we normalized the maximum error of estimate for $\widehat{\mathcal{U}}$, E, by dividing it to the standard deviation of the sampling distribution of $\widehat{\mathcal{U}}$. Column NE in Table 6 states that the maximum error of estimate for utilization is about 1.67 of standard deviation for all ISPs.

4 Related Work

Most of the studies in the Internet topology measurement field is based on collecting raw topology data from multiple vantage points by path sampling via

`traceroute` or `traceNET`. Then inferring routers and subnets using the raw data [7, 12, 10]. Early studies in the area used path sampling based topologies to derive conclusions about the topological characteristics of the Internet [3, 4]. Later on researchers argued about the validity of those claims by pointing out the limitations of path sampling due to source dependency and routing [8, 2].

`TraceNET` [14] infers subnets on an end-to-end path. Although `traceNET` substantially improves existing Internet topology maps by discovering subnet structures, it performs path sampling and hence suffers from above mentioned problems. `ExploreNET` [16] used in this paper allows us to discover individual subnets rather than subnets on an end-to-end path. In this study, we devised a sampling framework for studying network layer subnet characteristics in the Internet and introduced a set of unbiased and biased with small variance estimators for a set of subnet features. To the best of our knowledge, this is the first study investigating statistical sampling in the domain of subnet inference.

5 Conclusions

Given its large scale, drawing a complete *subnet level* topology of the Internet is neither practical nor economical. Consequently, sampling becomes a viable solution to derive the properties of subnets in the Internet. In this paper, we have proposed a framework for sampling subnets in the Internet along with either unbiased or slightly biased estimators with small variance for various subnet characteristics including total number of subnets, subnet prefix length distribution, mean subnet degree, and subnet IP address range utilization. Our theoretical and empirical evaluations on the estimators show that they work well in estimating population parameters.

References

1. The Internet Topology Zoo, available at http://www.topology-zoo.org/
2. Achlioptas, D., Clauset, A., Kempe, D., Moore, C.: On the bias of traceroute sampling. In: ACM STOC. Baltimore, MD, USA (May 2005)
3. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: SIGCOMM. New York, NY, USA (October 1999)
4. Govindan, R., Tangmunarunkit, H.: Heuristics for Internet map discovery. In: IEEE INFOCOM. Tel Aviv, Israel (Mar 2000)
5. Heidemann, J., Govindan, R., Papadopoulos, C., Bartlett, G., Bannister, J.: Census and survey of the visible internet. In: ACM IMC. Vouliag., Greece (Oct 2008)
6. Irwin Miller, M.M.: John E. Freund's Mathematical Statistics with Applications. Prentice Hall (October 2003)
7. K. Claffy, Tracie E. Monk, D.M.: Internet Tomography. Nature (Jan 1999)
8. Lakhina, A., Byers, J., Crovella, M., Xie, P.: Sampling biases in IP topology measurements. In: IEEE INFOCOM. San Francisco, CA, USA (Mar 2003)
9. Mann, P.S.: Introductory Statistics. Wiley (February 2010)
10. Shavitt, Y., Shir, E.: DIMES: Distributed Internet measurements and simulations, project page http://www.netdimes.org
11. Sherwood, R., Bender, A., Spring, N.: DisCarte: A disjunctive internet cartographer. In: ACM SIGCOMM. Seattle, WA, USA (Aug 2008)
12. Spring, N., Mahajan, R., Wetherall, D., Anderson, T.: Measuring ISP topologies with Rocketfuel. IEEE/ACM Transactions On Networking 12(1), 2–16 (Feb 2004)
13. Thompson, S.K.: Sampling. Wiley-Interscience (April 2002)
14. Tozal, M.E., Sarac, K.: TraceNET: An internet topology data collector. In: ACM Internet Measurement Conference. Melbourne, Australia (Nov 2010)
15. Tozal, M.E., Sarac, K.: Palmtree: An ip alias resolution algorithm with linear probing complexity. Computer Communications 34, 658–669 (April 2011)
16. Tozal, M.E., Sarac, K.: Subnet level network topology mapping. In: IEEE IPCCC. Orlando, FL, USA (November 2011)