Sampling Biases in IP Topology Measurements

Anukool Lakhina John W. Byers Mark Crovella Peng Xie
Department of Computer Science
Boston University
{anukool, byers, crovella, xp}@cs.bu.edu

***Abstract*— Considerable attention has been focused on the properties of graphs derived from Internet measurements. Router-level topologies collected via traceroute-like methods have led some to conclude that the router graph of the Internet is well modeled as a power-law random graph. In such a graph, the degree distribution of nodes follows a distribution with a power-law tail.**

We argue that the evidence to date for this conclusion is at best insufficient. We show that when graphs are sampled using traceroute-like methods, the resulting degree distribution can differ sharply from that of the underlying graph. For example, given a sparse Erdös-Rényi random graph, the subgraph formed by a collection of shortest paths from a small set of random sources to a larger set of random destinations can exhibit a degree distribution remarkably like a power-law.

We explore the reasons for how this effect arises, and show that in such a setting, edges are sampled in a highly biased manner. This insight allows us to formulate tests for determining when sampling bias is present. When we apply these tests to a number of well-known datasets, we find strong evidence for sampling bias.

I. Introduction

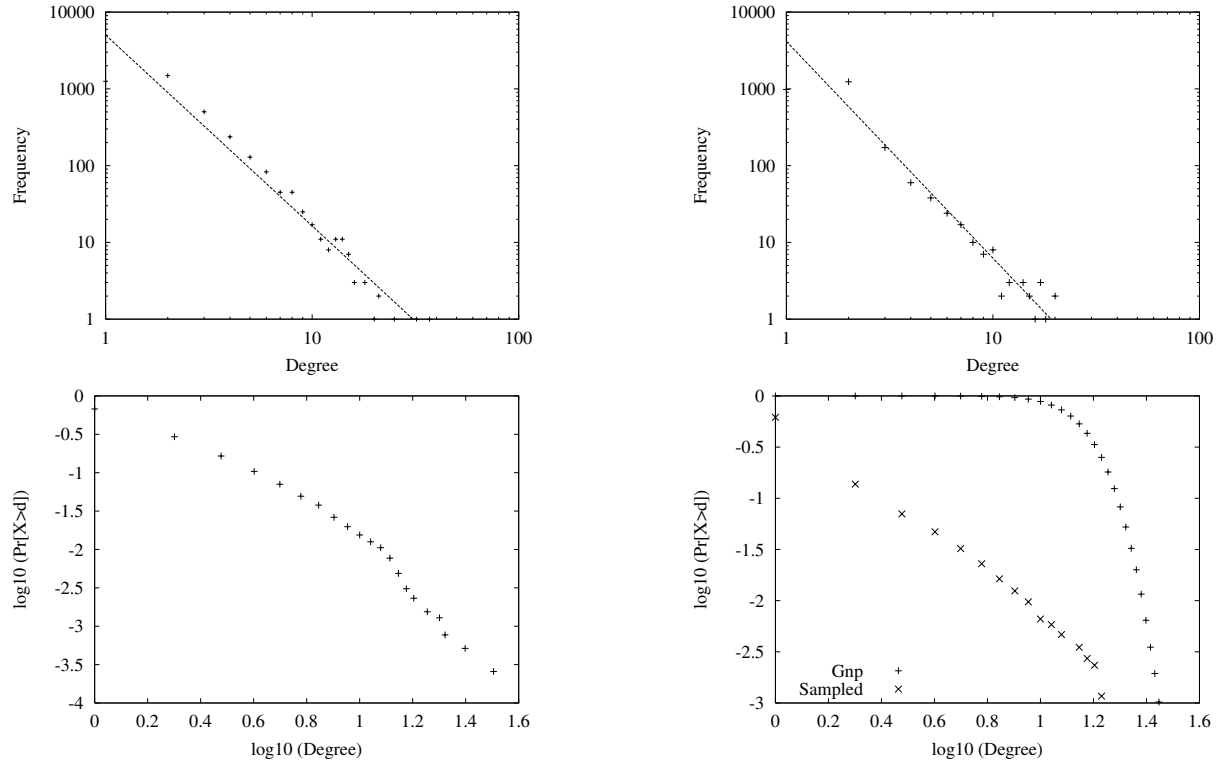
A significant challenge in formulating, testing and validating hypotheses about the Internet topology is a lack of highly accurate maps. This problem is especially acute when studying the router-level topology, the graph formed by taking the set of routers as vertices and adding an edge between any pair of routers which are one IP hop apart. In lieu of accurate maps, researchers currently rely on a variety of clever probing methods and heuristics to assemble an overall picture of the router-level topology. One such strategy is to use traceroute, a probing tool which reports the interfaces along the IP path from a source to a destination. Intuitively, traceroute has the capability to sample an end-to-end path through the network. If one assimilates the results of a large number of traceroutes, each of which sheds a small amount of light on the underlying connectivity of the router-level topology, the resulting sampled subgraph is a reflection of the entire topology. But how accurate a reflection does this procedure produce? It is well known that the process is not perfect [1]. For instance, it is only possible to run traceroute from a cooperating source machine, thus the choice of sources in such an experiment is highly constrained. Furthermore, some routers ignore traceroute probes, and others respond incorrectly. Nevertheless, these methods, or closely related methods, are widely used in mapping studies such as [2]–[6] and provide the basis for drawing deeper conclusions about the Internet topology as a whole [7]–[10].

One such conclusion, and indeed, one of the most surprising findings reported in [7], is evidence for a power-law relationship between frequency and degree in the router-level topology. Using their formalism, consider the router-level topology $G = (V, E)$ where vertices in V correspond to routers and undirected edges in E correspond to one hop IP connectivity between routers. Then, let d be a given degree, and define f_d to be the frequency of degree d vertices in G, *i.e.* $f_d = \#\{v \in V \text{ s.t. } \#\{(v, x) \in E\} = d\}$. The power-law relationship they then provide evidence for is $f_d \propto d^c$, for a constant power-law exponent c. At the time their study was conducted, maps of the router-level topology were scarce; one of the very few available was a dataset collected by Pansiot and Grad in 1995 [2]. The evidence for the frequency vs. degree power-law (reproduced directly from the dataset in [2]) is presented in Figure 1(a) as a plot on log-log scale. The upper graph is a plot of the pdf as it originally appeared in [7]; the lower graph is a plot of the log-log complementary distribution (ccdf).

As noted earlier, and as with other maps collected from traceroute-based methods, the Pansiot and Grad inventory of routers and links was undoubtedly incomplete. However, there is a more serious problem with drawing conclusions about characteristics of the router-level topology from this dataset (or any similar traceroute-driven study) than that of incomplete data, namely *sampling bias*.

In a typical traceroute-driven study [9], traceroute destinations are passive and plentiful, while active traceroute sources require deployment of dedicated measurement infrastructure, and are therefore scarce. As such, when traces are run from a relatively small set of sources to a much larger set of destinations, those nodes and links closest to the sources are sampled much more frequently than those that are distant from the sources and destinations. To demonstrate the significant impact this sampling bias can cause, we show the following experiment (more details and variations are in Section II).

We are interested in the subgraph induced by taking a sample of nodes and edges traversed by paths from k sources to m destinations, and focus on whether the measured degree distribution in the subgraph is representative of the entire graph. We choose $G = (V, E)$ to be a $G_{N,p}$ graph using the classical Erdös-Rényi graph model, *i.e.* where $|V| = N$ and where each edge (u, v) is chosen to be present in E independently with probability p. Modeling the intricacies of IP routing is beyond the scope of this experiment; we simply assign edges random weights $1 + \epsilon$, where ϵ is chosen



(a) Pansiot-Grad (b) Subgraph sampled from $G_{N,p}$

Fig. 1. Evidence for a Frequency Vs Degree Power Law in (a) the Pansiot-Grad dataset and (b) a sampled subgraph of a random graph.

uniformly at random from $[-\frac{1}{N}, \frac{1}{N}]$ and use shortest-path routing (the random weights are chosen solely to break ties between shortest-path routes).

In Figure 1(b), we present a frequency vs. degree plot on log-log scale of the induced subgraph when k = 1, m = 1000, N = 100,000, and Np = 15 (where Np is the average degree of a vertex). These parameters were chosen specifically to provide visual similarity to the plot from the dataset in [2]; we report on similar results for many other parameter settings later in the paper. While the induced subgraph demonstrates a similar frequency vs. degree power-law fit, this is a *measurement artifact* and is *not* representative of the underlying random graph. The degree distribution of the underlying random graph is far from a power-law; it is well-known to be Poisson. The degree distribution of the sampled graph is contrasted with the degree distribution of the underlying graph in the lower plot of Figure 1(b), which shows how different they are.

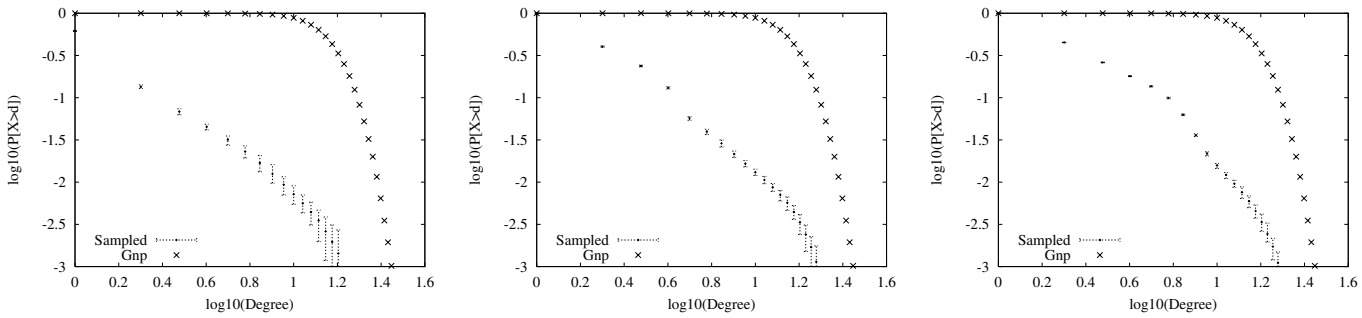
Clearly, this sort of misidentification is more likely when the data values only span a narrow range, as is the case here. In fact, the data used to argue for a power-law distribution in [7] spans a range of values that is too small for conclusive judgements. However, even over this narrow range, the difference between the two distributions in Figure 1(b) is so great as to be important for modeling purposes.

These observations form the motivation for our work and lead us to the following questions which we will study in this paper. What are the root causes of sampling bias in traceroute mapping studies? Are observed power-laws in router degree distributions a fact or a measurement artifact? Can we detect sampling bias in well known traceroute datasets?

We explore the sources and effects of sampling bias in several stages. First, in Section II, we investigate sampled subgraphs on *generated topologies*, namely classical random graphs and power-law random graphs (PLRGs), and expand upon and develop the arguments presented earlier in the introduction. We then explore possible sources of sampling bias. Next, we analytically examine the causes for sampling biases in Section III and formulate tests to detect the presence of sampling bias. Then, in Section IV, we consider a number of traceroute-based datasets, and conclude that they show evidence of sampling bias.

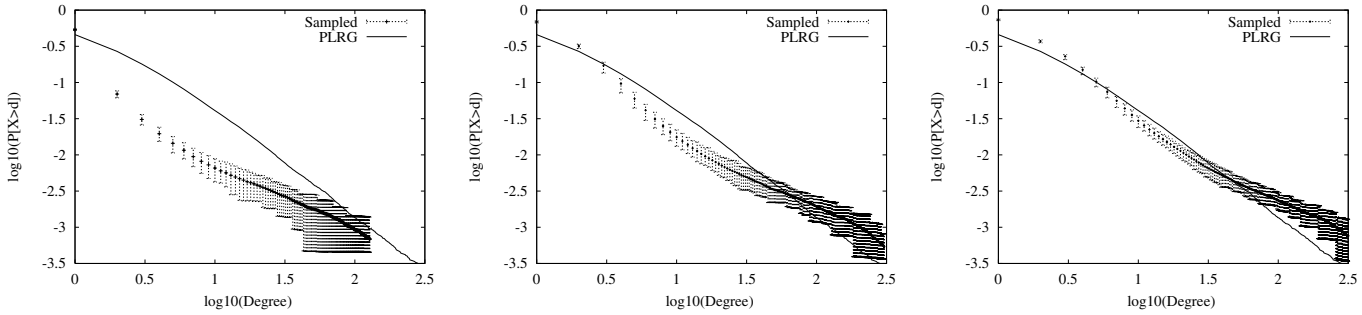
II. Examining Node Degree Distribution of Sampled Subgraphs

The previous section showed an example of a sampled subgraph whose degree distribution deviates substantially from the degree distribution of the underlying topology. In this section, we present further evidence of a prevalent sampling bias across a broad spectrum of sampled subgraphs on both classical random graphs [11] and power-law random graphs derived from the PLRG model [12]. We then examine possible sources of the bias responsible for the disparity between the degree distribution of the underlying graph and the degree distribution of the sampled graph.



(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 2. Degree Distribution of subgraph sampled from Erdös-Rényi random graph ($N = 100,000$, $p = 0.00015$)



(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 3. Degree Distribution of subgraph sampled from power-law random graph ($N = 112,959$, $\eta = 2.1$)

We begin by introducing our experimental setup, relevant terminology and assumptions.

A. Definitions and Assumptions

Let $G = (V, E)$ be a given sparse undirected graph with $|V| = N$. Our experimental methodology assigns random real-valued weights to the edges as follows: for all edges e in E, let the link weight $w(e) = 1 + \epsilon$ where ϵ is chosen uniformly and independently for each edge from the interval $[-\frac{1}{N}, \frac{1}{N}]$. Then assume that we have k distinct source vertices selected at random from V, and m distinct destination vertices also selected at random. For each source-destination pair, we compute the shortest path between them. We let $\hat{G}$ denote the graph (edges and vertices) induced by taking the union of the set of shortest paths between the k sources and m destinations. We will often refer to G as the *underlying* graph and $\hat{G}$ as the *sampled* graph. We will call such an experiment a *(k,m)-traceroute study*.

This experimental setup aims to model the prevalent methodology employed to discover the Internet topology. In a typical traceroute-driven study, point-to-point measurements conducted from a set of distributed vantage points to a large set of destinations are used to shed light on the underlying topology. Of course, our simple model does not attempt to capture all of the intricacies that such a live study encounters, *i.e.,* the complexities of IP routing, BGP policies, and topological location of end-points. But it is sufficient to explose potential sources of bias.

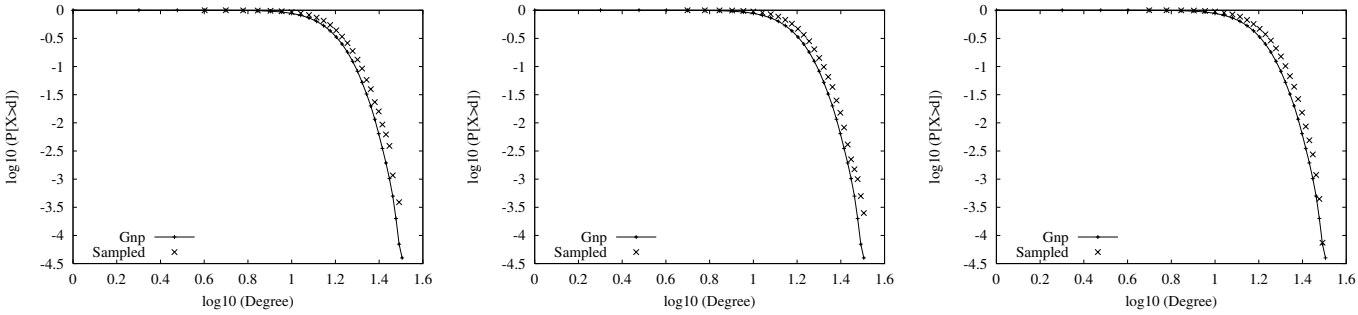
We begin by presenting experimental evidence of measurement bias in two choices of underlying graphs: graphs generated by the classical Erdös-Rényi random graph model [11] and graphs generated by the power-law random graph (PLRG) model [12]. These two graph models can be thought of as lying at two extremes of the degree spectrum: the degree distribution of classical random graphs is Poisson, while the degree distribution of PLRG graphs follows a power-law.

B. Sampling Random Graphs

Our first set of experiments employ the classical Erdös-Rényi random graph model described in the introduction. In all the random graphs we consider the average degree, Np, is sufficiently large so that the graph is connected with high probability. For our experiments, we ensured that each generated graph was connected.

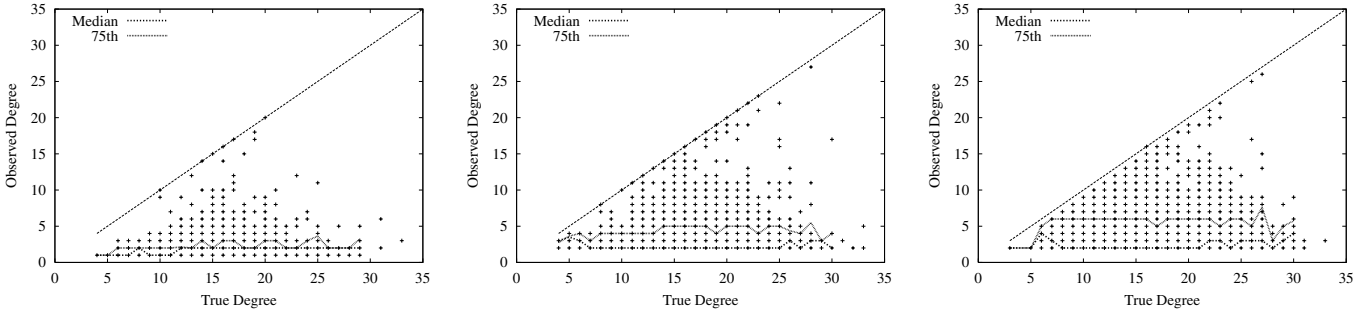
Figure 2 shows the degree distribution of $\hat{G}$ induced by $k = 1, 5, 10$ sources and $m = 1000$ destinations. Our underlying graph in this case has 100,000 nodes and 749,678 edges ($p = 0.00015$) with average degree 15. Each plot shows the 90% confidence intervals of 100 trials.

The results presented in these plots are important for two reasons. First, the degree distribution of $\hat{G}$, while not a strict power-law, is clearly long-tailed in each instance and can be potentially mistaken for (or approximated by) a power-law. Second, the degree distribution of $\hat{G}$ is vastly different from the true Poisson degree distribution of $G_{N,p}$, implying that $\hat{G}$ is not a representative sample of our underlying G. As



(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 4. *Conjecture 1:* Are the nodes of $G_{N,p}$ sampled disproportionately? These graphs show the true degree distribution for nodes in $\hat{G}$ along with the complete degree distribution of the underlying $G_{N,p}$ (as CCDF on log-log axes). Since both degree distributions are similar, we discard this conjecture.



(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 5. *Conjecture 2:* Are the edges of $G_{N,p}$ sampled disproportionately? These graphs show that the number of edges discovered incident to a node in $\hat{G}$ is not proportional to the node's true degree.

such, conclusions that are made about $\hat{G}$ (*e.g.,* explanations put forward to explain the measured degree distribution ([13], [14]) may not necessarily apply to the underlying G.

C. Sampling Power-Law Graphs

Since sampled subgraphs of random graphs yield highly variable degree distributions, it is natural to wonder what sampled subgraphs of power-law graphs yield.

Our second set of experiments repeat similar shortest path simulations on the power law random graph model of [12] (PLRG). Briefly, given N nodes and exponent η, the PLRG model initially assigns degrees drawn from a power-law distribution with exponent η and then proceeds to interconnect the nodes as follows. First, each node v with target degree d_v is replaced with d_v nodes; then the resulting set of nodes in the graph are connected by a random matching. The copied nodes and their incident edges are then collapsed. After removing self-loops and multi-edges, the largest connected component is then extracted. We used this procedure to construct a power-law graph with $112,959$ nodes, $186,629$ edges, and power-law exponent of about 2.1 (to match the exponent found by [7] in their router dataset).

Figure 3 shows the degree distribution of $\hat{G}$ induced by 1, 5, and 10 sources to 1000 destinations. Here, the sampled graph $\hat{G}$ exhibits a degree distribution visually similar to the underlying G. Further, a handful of sources are sufficient to produce a degree distribution similar to that of the underlying PLRG graph. This is in contrast to our earlier experiments on $G_{N,p}$, where we found that the sampled $\hat{G}$ exhibited a clearly distinct distribution. To understand this difference better, we now investigate possible sources of bias when sampling $G_{N,p}$.

D. Sources of Sampling Bias

We have presented evidence demonstrating that the sampled graph $\hat{G}$ can be vastly different from the underlying graph G. We now attempt to identify *where* biases arise when sampling $G_{N,p}$ graphs. Our explanations stem from observations of extensive simulations; we subsequently present an analytical justification.

An initial conjecture to explain this phenomena is that nodes are sampled disproportionately. Certainly $\hat{G}$ has fewer nodes, but are these nodes a uniform sample from the nodes in the underlying graph? This posits that shortest path routing favors the higher degree nodes of $G_{N,p}$ in the computed paths. In such a scenario, high degree nodes of $G_{N,p}$ reduce the distance to reach destination nodes and so become frequently explored intermediate nodes. To explore this conjecture, we study the true degree distribution of nodes in $\hat{G}$, *i.e.,* for each node n in $\hat{G}$, we examine how many neighbors n has in the underlying $G_{N,p}$ graph.

Figure 4 plots the true degree distribution for nodes in various instances of $\hat{G}$ along with the degree distribution of nodes in $G_{N,p}$. Contrary to our intuition, the true degree

distribution of $\hat{G}$ is similar to the degree distribution of nodes in $G_{N,p}$. Therefore, it is not the selection of nodes by shortest path routing that is biased.

A natural second conjecture for the source of sampling bias concerns edges. One consequence of taking measurements using a small number of sources and relying on an end-to-end strategy, is that edges are selected disproportionately. Clearly $\hat{G}$ has fewer edges than the underlying $G_{N,p}$. But are the number of edges discovered incident to a node proportional to its true degree? To explore this conjecture, we compare the number of edges discovered for each node in $\hat{G}$ with its true degree.

This comparison is shown in Figure 5, where each node's true degree is plotted against its observed degree in $\hat{G}$. The dotted line $y = x$ corresponds to observing the true degree. In each of these plots, lines corresponding to the median and 75th percentile of observed degrees for a given (true) degree are also shown. If edges are selected uniformly, then the ratio of observed edges to true degree of a node should be constant. Thus, if this were indeed the case, we should expect to see points tightly clustered around a trend line $y = cx$ for some $c < 1$. Instead, Figure 5 shows no such trend. In fact, the number of edges observed incident to a high degree node is comparable to that of an average degree node. These plots therefore support our second conjecture: bias arises when edges incident to a node in the underlying graph are sampled disproportionately. In the next section, we analytically explore the reasons for this effect.

III. Analysis and Inference

Now we seek to understand the nature of sampling bias via analysis; using this understanding we then develop criteria for detecting the presence of sampling bias in empirical data.

A. Analyzing Sampling Bias

The previous sections have shown that an important source of sampling bias in the experiments described here is the failure to observe edges which exist but are not part of the shortest-path trees.

To explore the nature of this kind of sampling bias, we turn to analysis. In this section we concern ourselves only with the single-source shortest path tree ($k = 1$). We are concerned with the visibility of edges provided by this tree, so the particular question we ask is: *Given some vertex in $\hat{G}$ that is h hops from the source, what fraction of its true edges (those in G) are contained in the subtree ($\hat{G}$)?* That is, how does visibility of edges decline with distance from the source?

Our analysis assumes $G_{N,p}$ graphs like those defined in Section II-A. Let the number of destinations be m, the number of vertices in G be N, and the probability that two vertices in G are connected be p. In this case we can state the following result.

Theorem 1: Let $p_h(n)$ denote the probability that the shortest path to n destinations ($n \leq m$) passes through a given edge of a given vertex at h hops from the source. Then:

$$p_h(n) = \sum_{j=0}^{\infty} P(Np, j) \sum_{k=0}^{m} p_{h-1}(k) \sum_{i=0}^{k} B(k, |\Gamma_h|/N, i) B(k-i, 1/j, n) \quad \text{for } h > 0,\ n = 0, \ldots, m$$

and

$$p_0(n) = \sum_{j=0}^{\infty} P(Np, j) \sum_{i=0}^{m} B(m, 1/N, i) B(m-i, 1/j, n) \quad \text{for } n = 0, ..., m$$

where $B(n, p, x)$ denotes the Binomial distribution, stating the probability of x successes in n trials each having success probability p; $P(\lambda, j)$ is the Poisson distribution, used here to describe the probability of a vertex having j edges in a random graph with average degree λ; and Γ_h denotes the set of vertices in G at distance h from the source.
For the proof of Theorem 1 see [15].

In order to evaluate this expression we need $|\Gamma_h|$. In [16], a number of bounds are given for $|\Gamma_h|$, and similar results are developed in [17]; however in general, tight bounds for this expression over the entire range of h are not known. As a result we use an approximation to $|\Gamma_h|$ derived from simulation and consistent with the bounds in [16], [17].

Using Theorem 1, we can study how visibility of edges declines with distance from the source. The probability that an edge in G that is connected to a vertex in $\hat{G}$ is actually observed (*i.e.,* is part of $\hat{G}$) is $1 - p_h(0)$. (This excludes the edge connecting the vertex to its parent in the tree.) This probability tells us how biased our node degree measurements become as a function of distance from the source. When this probability is small, we are missing most edges and so our estimates of node degree will be very inaccurate.

In Figure 6 we plot this value as a function of h (the distance from the source node). In each case, $Np = 15$ and we vary the number of destinations m from 100 to 1000. We show two cases to illustrate different experimental situations. On the left the number of vertices in G is 10,000; this value is chosen so that the number of destinations encompasses a non-negligible fraction of G. On the right the number of vertices in G is 1,000,000; in this case, the number of destinations is very small compared to the size of G.

The plots show that over the vast majority of nodes in $\hat{G}$, visibility of edges is quite poor. Only at hops 0 (the source) and 1 are a majority of edges discovered; and for hop 1, a large fraction of edges are not discovered unless the number of destinations is large. Comparing Figures 6(a) and (b), we can see that the number of nodes in the underlying graph does not have a strong effect on visibility; regardless of the size of

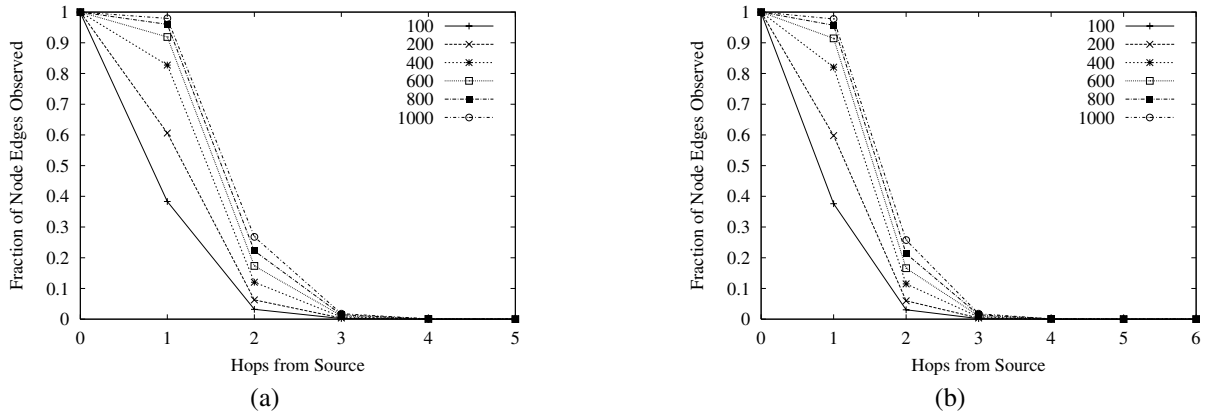


Fig. 6. Visibility of Edges with Varying Number of Destinations (average degree=15); (a) N = 10,000; (b) N = 1,000,000.

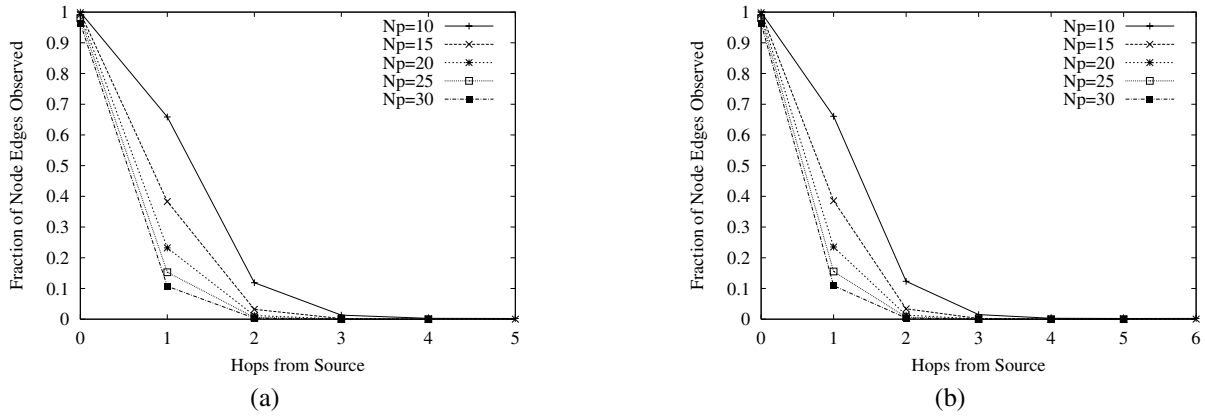


Fig. 7. Visibility of Edges with Varying Vertex Degree in G (number of destinations=100); (a) N = 10,000; (b) N = 1,000,000.

G, visibility of edges is essentially restricted to one or two hops from the source.

To further explore how the limits of visibility depend on the properties of the underlying graph G, we consider the effects of varying the average degree of a vertex (Np). The results are shown in Figure 7, for 100 destinations. The figure shows that when vertex degree is small, visibility is extended slightly. However the sharp decline in visibility remains even at relatively low vertex degree.

These results show that shortest-path trees only effectively explore a very small neighborhood around the source in a random graph. This helps explain the effect observed in Figure 5. Furthermore, these results suggest that the degree distribution observed close to the source may be quite different from the distribution observed far from the source; in the next subsection we develop this idea further and use it to examine graphs derived from traceroute measurements.

B. Inferring the Presence of Bias

In the previous subsections, we provided evidence for and identified sources of bias in (k,m)-traceroute studies. Given these findings, a natural question to ask is if it is possible to detect evidence of bias in similar measurements when the underlying topology is unknown.

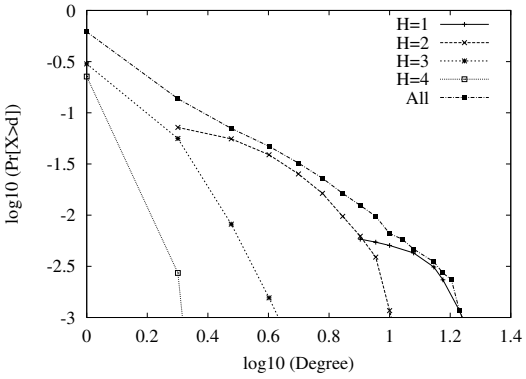
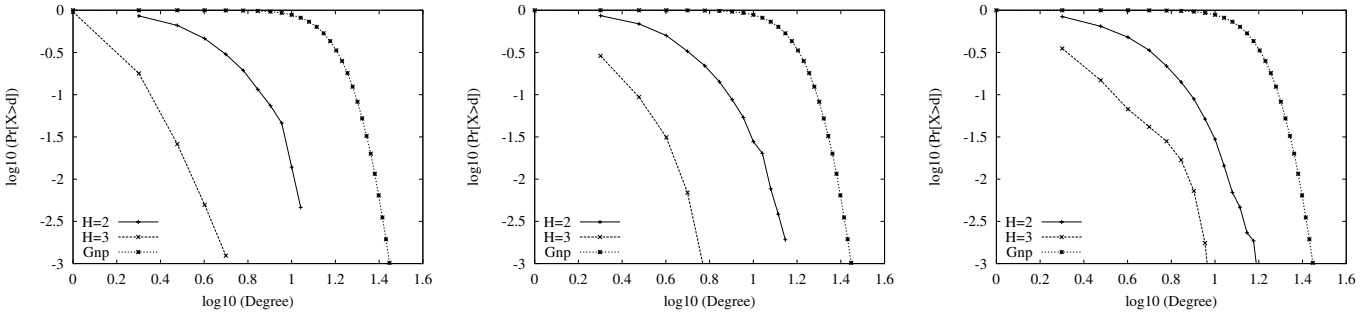


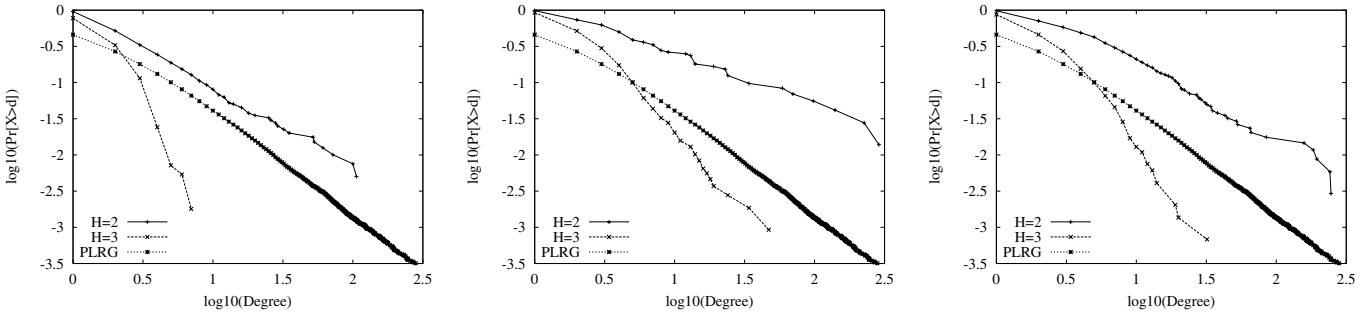
Fig. 8. $\Pr[D \cap H]$ for subgraph sampled from $G_{N,p}$ (1 source, 1000 destinations)

We start from the observation made in the last subsection, which showed that nodes close to the measurement source were explored more completely than those further from the source. This suggests that conditioning our measurements on distance from the source may be fruitful. Our general idea is that if measurements are unbiased, then their statistical properties should not change with distance from the source.



(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 9. Degree distribution by hop distance from source(s), $\Pr[D|H]$, for subgraphs sampled from $G_{N,p}$



(a) 3 sources, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Fig. 10. Degree distribution by hop distance from source(s), $\Pr[D|H]$, for subgraphs sampled from PLRG.

However, if measurements are biased, we should be able to detect that by looking at statistics as a function of distance from the source.

To explore this idea, we study the conditional probability that a node has degree d given that it is at hop h from the source. With $k > 1$ sources, we define h as the minimum hop distance over all sources. We denote this conditional probability by $\Pr[D|H]$.

Multiplying $\Pr[D|H]$ by $\Pr[H]$ yields the joint probability $\Pr[D \cap H]$, which is shown in Figure 8 for a subgraph sampled from a $G_{N,p}$ graph by 1 source and 1000 destinations. This figure shows how node degrees at each hop together produce the illusion of an overall long-tailed degree distribution of $\hat{G}$. Figure 9 shows the hop conditioned degree distribution, $\Pr[D|H]$, for sampled subgraphs of $G_{N,p}$. We make two observations from Figures 8 and 9. First, the highest degree nodes are found at small h, that is, at hops nearest to the source nodes. Second, the overall distributions $\Pr[D|H]$ vary considerably with h. These two observations suggest criteria for detecting bias in (k,m)-traceroute studies:

- **C1** *Do the highest-degree nodes tend to be near the source(s)?* If so, this is consistent with bias, since in an unbiased sample the highest-degree nodes should be randomly scattered throughout $\hat{G}$.
- **C2** *Is the distributional shape near the source different from that further from the source?* Again, if so, this is consistent with bias, since this property should not vary within an unbiased sample.

If these criteria are to be useful they should hold for the case of biased sampling of other random graphs as well. Figure 10 shows $\Pr[D|H]$ for graphs sampled from PLRG graphs. First, we see that the highest degree nodes are generally at hop 2 — close to the source. Next, when $\Pr[D|H]$ is compared for different values of h, we see that there are sharp differences between cases for h=2 and h=3.

The consistent behavior of the criteria **C1** and **C2** on samples from both $G_{N,p}$ and PLRG graphs suggests that they can help identify cases in which measurements taken from unknown underlying graphs may be subject to bias. This leads us to a formalization of *bias* in this setting, which we can define as a failure of a sampled graph to meet statistical tests for randomness associated with these two criteria. A dataset failing to show randomness under both criteria would seem to be a poor choice for use in making generalizations about the true nature of the underlying graph.

C. Formal Tests for Criteria

To quantitatively evaluate criteria **C1** and **C2**, we now develop formal tests.

Let m be the median hop distance of a node in the vertex set from the source. We partition the vertex set into two subsets, $\mathcal{N}$ and $\mathcal{F}$ (near and far), where $\mathcal{N}$ consists of those vertices of hop distance strictly less than m from the source, and $\mathcal{F}$ consists of those vertices of distance at least m from the source. We then let $p_{\mathcal{N}}$ denote the fraction of all vertices in $\mathcal{N}$ (typically slightly less than 0.5).

To test criterion **C1**, we ask whether the 1% highest-degree vertices tend to appear unusually often in $\mathcal{N}$ as oppposed to $\mathcal{F}$. More precisely, we take the $|V|/100$ vertices with largest rank (plus nodes whose degree ties that of the node with rank $|V|/100$) and determine the number of these vertices which lie in $\mathcal{N}$. Suppose the number of vertices thus considered is ν, and out of those ν the number happening to lie in $\mathcal{N}$ is κ. An outcome in which κ deviates substantially from the expected value ($p_{\mathcal{N}}\nu$) is improbable. More formally, by Chernoff bounds [18, Ch. 4], we can bound the likelihood of an outcome which exceeds the mean by at least a $(1+\delta)$ multiplicative factor[1] by:

$$\Pr[\kappa > (1+\delta)\, p_{\mathcal{N}}\nu] < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{p_{\mathcal{N}}\nu}$$

We formulate the null hypothesis associated with **C1** as: ($\mathcal{H}_0^{C1}$) The 1% highest-degree nodes occur at random with respect to distance from the source(s). We can then reject the null hypothesis $\mathcal{H}_0^{C1}$ with confidence $1-\alpha$ when:

$$\alpha \geq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{p_{\mathcal{N}}\nu}$$

To test criterion **C2,** we first note that that the set of vertices V is the union of $\mathcal{N}$ and $\mathcal{F}$. Therefore, to test whether the degree distribution of the nodes in $\mathcal{N}$ differs from that in $\mathcal{F}$, it is sufficient to determine whether the degree distribution of nodes in $\mathcal{N}$ differs from those in V.[2]

The standard method for testing whether a observed dataset is likely to have been drawn from a particular distribution is the chi-square goodness-of-fit test [19, Ch. 27]. Given a set of observed data and an expected distribution, a histogram with ℓ bins is prepared from the observed data. The test statistic is then defined as:

$$\chi^2 = \sum_{i=1}^{\ell} (O_i - E_i)^2 / E_i$$

where O_i and E_i are the observed and expected frequencies respectively. The null hypothesis (the data is drawn from the given distribution) can be rejected with confidence level α if the computed χ^2 is greater than $\chi^2_{[1-\alpha;\ell-1]}$, obtained from the $\chi^2_{1-\alpha}$ distribution. The validity of the chi-square test statistic is sensitive to how the data is binned; to ensure validity we chose bin sizes keeping E_i is sufficiently large. To use this test, we form the following null hypothesis: ($\mathcal{H}_0^{C2}$) The degree distributions of datasets $\mathcal{N}$ and $\mathcal{F}$ are consistent with having been drawn randomly from V. When employing this test we use $\alpha = 0.005$ (corresponding to 99.5% confidence).

In summary, we define bias in (k,m)-traceroute studies as the rejection of null hypotheses $\mathcal{H}_0^{C1}$ and $\mathcal{H}_0^{C2}$ at a high confidence level (99.5% or higher). In the next section we use these tests to explore three well-known datasets.

[1]The reader may note that the ν trials are not fully independent. However, the weak dependence between each of the trials is in our favor, *i.e.* the presence of one node from the set of ν in $\mathcal{N}$ *decreases* the likelihood that a second node is also in $\mathcal{N}$, and thus the Chernoff bound is somewhat conservative.

[2]The complementary test (whether the degree distribution of $\mathcal{F}$ differs from V) will yield the same statistic because of the nature of the calculation.

IV. Examining Node Degree Distribution of Traceroute Datasets

Having defined tests to detect bias, we now turn to examining existing IP topology measurements.

A. From Models to Datasets

Before turning to empirical data, it is helpful to assess the ways in which real data differs from the idealized (k,m)-traceroute studies we have considered so far.

An example of the state-of-the art in topology measurement is CAIDA's Skitter project [4], which consists of roughly a dozen measurement monitors sending traceroute-like probes to a predetermined set of destinations. The differences between our experiments and a system like Skitter are at least twofold. First, we have assumed that sources and destinations are randomly placed in the graph. In a real measurement system, the location of sources in particular is constrained by the mechanics of setting up active measurement sites. It is possible that the neighborhoods around sources are unusual for this reason; while this introduces a new and different form of bias it would nevertheless be detected by our tests. Second, we have assumed that routing follows shortest paths, rather than paths dictated by a combination of IGP and EGP policies. While such an assumption has been made elsewhere [20], [21], it does not reflect the inflating effect that routing policy has on paths in the Internet [22], [23]. However routing policy is designed in general to find *short* paths, and the kinds of sampling bias we consider here would seem to be present in any system trying to keep paths short.

B. Datasets

We use three different snapshots of the router topology collected at different time periods.[3] Table I summarizes these datasets. Our first dataset, *Pansiot-Grad* routers, dates from 1995 [2]. This dataset was first used as evidence for power-law router degree distribution in the paper by Faloutsos *et al.* [7]. The next dataset is *Mercator,* collected subsequently to and much larger than the Pansiot-Grad set [3]. The authors of [3] also found evidence for a power-law degree distribution in this dataset.[4] The third dataset was obtained subsequent to Mercator, from 8 distinct sources of the *Skitter* project (after resolving interfaces to routers). It too shows evidence of a long-tailed degree distribution, as discussed in [8].

For the Mercator dataset, the hop distance from the source we use is computed by a shortest paths algorithm. This is not entirely accurate as it does not capture the measured path that the Mercator probe packets took. However, better path

[3]Perhaps the largest IP topology snapshots are recent measurements obtained by Skitter, *e.g.,* [9]. Unfortunately these datasets do not resolve interfaces to routers, and so introduce another complication in trying to assess router degree distribution.

[4]To be precise, the authors of [3] concluded that while the degree distribution upto a degree of 30 displayed evidence for a power-law, the distribution of higher degrees was more diffused.

information is not available for this dataset. For all other datasets, we have IP path information and rely on it to compute hop distance.

Dataset Name	Date	# of Nodes	# of Links
Pansiot-Grad	1995	3,888	4,857
Mercator Routers	1999	228,263	320,149
Skitter Routers	2000	7,202	11,575

TABLE I
SUMMARY OF DATASETS EXAMINED

C. Detecting Bias

Dataset	$\|V\|$	$p_{\mathcal{N}}$	ν	κ	Chernoff Bound	$\mathcal{H}_0^{C1}$
Pansiot-Grad	3,888	0.44	41	38	2×10^{-4}	Reject
Mercator Routers	228,263	0.45	2,290	2,065	10^{-172}	Reject
Skitter Routers	7,202	0.44	104	87	9×10^{-7}	Reject

TABLE II
TESTS OF HYPOTHESIS $\mathcal{H}_0^{C1}$

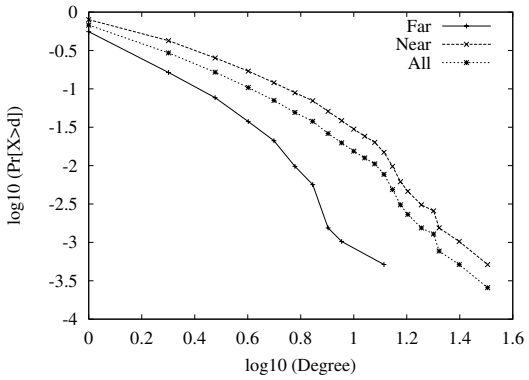
Dataset	ℓ	α	$\chi^2_{[1-\alpha;\ell-1]}$	χ^2	$\mathcal{H}_0^{C2}$
Pansiot-Grad	17	0.005	35.72	1082.0	Reject
Mercator Routers	123	0.005	167.4	59729	Reject
Skitter Routers	19	0.005	23.59	1965	Reject

TABLE III
TESTS OF HYPOTHESIS $\mathcal{H}_0^{C2}$

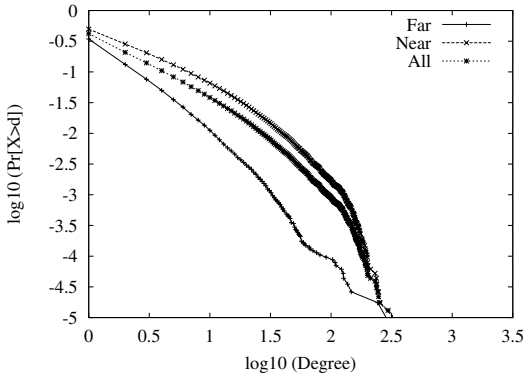
Table II summarizes the results of tests of the hypothesis $\mathcal{H}_0^{C1}$. The table shows that all three datasets appear to show bias under criterion **C1** with (much) greater than 99.5% confidence. Similarly, Table III summarizes the results of tests of the hypothesis $\mathcal{H}_0^{C2}$. The table shows that all three datasets appear to show bias under criterion **C2** as well at the 99.5% confidence level.

The difference between the distribution of nodes in $\mathcal{N}$, in $\mathcal{F}$, and in V is shown for the three datasets in Figure 11. These differences are the reason for the large χ^2 statistics in Table III. The figure agrees with the results of the statistical tests, namely that all three distributions are visually distinct for each dataset. Furthermore, the set of vertices in $\mathcal{N}$ tends to show higher median and average degree than the set of vertices in $\mathcal{F}$, which is consistent with the results of the **C1** tests as well.

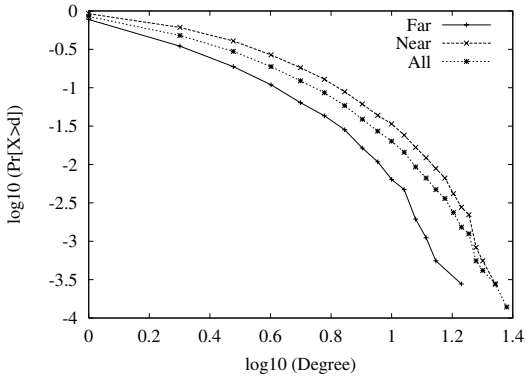
In summary, all three datasets pass our statistical tests for evidence of sampling bias; in each and every case we can reject the null hypothesis. Thus we can reasonably assume that the true degree distribution of Internet router-level graphs is different than that of any of these datasets. In particular, these tests suggest that the true router graph may have a higher proportion of high-degree nodes than would appear from simple extrapolation of these measurements.



(a) Pansiot-Grad Routers



(b) Mercator Routers



(c) Skitter Routers

Fig. 11. Examining $\Pr[D|H]$ for empirical datasets.

V. CONCLUSIONS

Drawing conclusions about the Internet topology from a set of distributed measurements, such as those collected in a traceroute-driven study, has long been known to be an imperfect process. The conventional wisdom is that collected measurement data is typically incomplete, noisy, and may not be representative. In this work, we have demonstrated the effects of a potentially much more serious flaw than that of noisy data: that of a pervasive bias in the topology data gathered by a traceroute-driven approach. On generated topologies, we demonstrate that the sampled subgraphs induced by a collection of source-destination shortest paths can have degree distributions which bear little resemblance to

those of the underlying graph. We present analytical support for this finding, as well as methods to test whether the properties of a measured subgraph show evidence of sampling bias. Applying these methods to three empirically measured router inventories shows strong evidence of sampling bias.

Our results suggest that since long-tailed degree distributions can arise simply through biased sampling of graphs, node degree distribution alone may not be a sufficiently robust metric for characterizing [7] or comparing router-level topologies [24], [25]

An interesting, and seemingly very difficult open question related to our work is that of conducting statistically unbiased random samples of properties of nodes and links in the Internet. Measurement methods targeted at a specific region of the Internet, such as those used by Rocketfuel to map ISP networks [5], have exploited the flexibility of selecting their end-points in an informed manner. These methods avoid some pitfalls of (k,m)-traceroute studies, and so are an attractive limited-scale alternative in light of the sampling bias effects demonstrated in our work. More generally, a technique with the capability to accurately sample the degree of a randomly chosen router in the Internet would be a useful tool in ascertaining the true degree distribution of the underlying network.

ACKNOWLEDGEMENTS

We gratefully acknowledge the anonymous referees for their feedback. This work was supported in part by NSF grants ANI-9986397 and ANI-0093296.

REFERENCES

[1] L. Amini, A. Shaikh, and H. Schulzrinne, "Issues with Inferring Internet Topological Attributes," in *Proceedings of SPIE ITCom*, 2002.

[2] J. Pansiot and D. Grad, "On Routes and Multicast Trees in the Internet," *ACM Computer Communication Review*, vol. 28, no. 1, pp. 41–50, January 1998.

[3] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet Map Discovery," in *Proceedings of IEEE INFOCOM*, March 2000.

[4] "Skitter," At *http://www.caida.org/tools/measurement/skitter*.

[5] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP Topologies with Rocketfuel," in *Proceedings of ACM SIGCOMM*, August 2002.

[6] N. Spring, D. Wetherall, and T. Anderson, "Scriptroute: A Public Internet Measurement Facility," in *4th USENIX Symposium on Internet Technologies and Systems*, 2002.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-Law Relationships of the Internet Topology," in *ACM SIGCOMM*, Cambridge, MA, September 1999, pp. 251–62.

[8] P. Barford, A. Bestavros, J. Byers, and M. Crovella, "On the Marginal Utility of Network Topology Measurements," in *Proc. of the SIGCOMM Internet Measurement Workshop*, November 2001.

[9] A. Broido and K. Claffy, "Connectivity of IP Graphs," in *Proceedings of SPIE ITCom*, August 2001.

[10] A. Lakhina, J. W. Byers, M. Crovella, and I. Matta, "On the Geographic Location of Internet Resources," in *Proc. of the SIGCOMM Internet Measurement Workshop*, November 2002.

[11] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[12] W. Aiello, F. Chung, and L. Lu, "A Random Graph Model for Massive Graphs," in *32nd Annual Symposium in Theory of Computing*, 2000.

[13] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, pp. 509–512, October 1999.

[14] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou, "Heuristically Optimized Tradeoffs," in *Proceedings of ICALP*, 2002.

[15] A. Lakhina, J. Byers, M. Crovella, and P. Xie, "Sampling Biases in IP Topology Measurements," Boston University Computer Science, Tech. Rep. BUCS-TR-2002-021, July 2002.

[16] F. Chung and L. Lu, "The diameter of random sparse graphs," *Advances in Applied Math*, pp. 257–279, 2001.

[17] P. van Mieghem, G. Hooghiemstra, and R. W. van der Hofstad, "A scaling law for the hopcount in the Internet," Delft University of Technology, Tech. Rep. Report 2000125, 2000.

[18] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge University Press, 1995.

[19] R. Jain, *The Art of Computer Systems Performance Analysis*. Wiley and Sons, Inc., 1991.

[20] P. van Mieghem, G. Hooghiemstra, and R. W. van der Hofstad, "On the Efficiency of Multicast," *IEEE/ACM Transactions on Networking*, May 2001.

[21] P. van Mieghem and M. Janic, "Stability of a Multicast Tree," in *Proceedings of IEEE INFOCOM*, 2002.

[22] H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin, "The Impact of Routing Policy on Internet Paths," in *Proceedings of IEEE INFOCOM*, 2001.

[23] H. Tangmunarunkit, R. Govindan, and S. Shenker, "Internet Path Inflation Due to Policy Routing," in *SPIE ITCom*, 2001.

[24] D. Magoni and J. Pansiot, "Analysis and Comparison of Internet Topology Generators," in *2nd International IFIP-TC6 Networking Conference*, 2002.

[25] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network Topology Generators: Degree-Based vs. Structural," in *Proceedings of ACM SIGCOMM'02*, Pittsburgh, PA, August 2002.