Sampling Biases in IP Topology Measurements*

BUCS-TR-2002-021

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

July 15, 2002

Abstract

Considerable attention has been focused on the properties of graphs derived from Internet measurements. Router-level topologies collected via traceroute studies have led some authors to conclude that the router graph of the Internet is a scale-free graph, or more generally a power-law random graph. In such a graph, the degree distribution of nodes follows a distribution with a power-law tail.

In this paper we argue that the evidence to date for this conclusion is at best insufficient. We show that graphs appearing to have power-law degree distributions can arise surprisingly easily, when sampling graphs whose true degree distribution is not at all like a power-law. For example, given a classical Erdös-Rényi sparse, 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 easily appear to show 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 distinguish measurements taken from the Erdös-Rényi graphs from those taken from power-law random graphs. When we apply this distinction to a number of well-known datasets, we find that the evidence for sampling bias in these datasets is strong.

*Supported in part by NSF grants ANI-9986397, ANI-0095988, and ANI-0093296.

1 Introduction

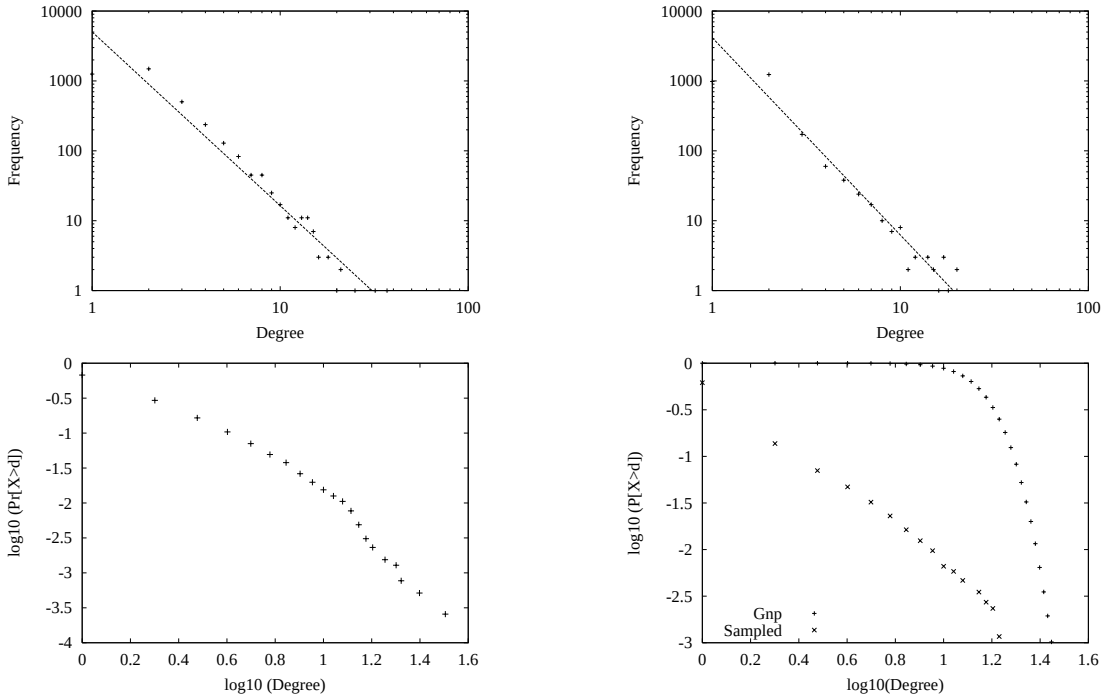
A significant challenge in formulating, testing and validating hypotheses about the Internet topology is a lack of highly accurate maps, a problem which is especially acute when studying the router-level topology. As such, researchers currently rely on a variety of clever probing methods and heuristics to assemble an overall picture of the network. One such strategy is the use of `traceroute`, a probing tool which reports the interfaces along the IP path from a source to a destination. By assimilating 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 mosaic is a reflection of the entire topology. But does this procedure result in an accurate reflection? Certainly there are limitations – some routers do not respond to traceroute probes; one must somehow gain confidence that the probes conducted provide sufficient and equal coverage across the entire Internet; and undoubtedly some nodes and links may not be reachable due to issues such as BGP policies. Nevertheless, these methods, or closely related methods, are widely used in mapping studies such as [11, 9, 8, 12] and provide the basis for drawing deeper conclusions about the Internet topology as a whole [7, 2, 3].

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 physical links 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 data-set collected by Pansiot and Grad in 1995 [11]. The compelling evidence for the frequency vs. degree power-law (reproduced directly from the dataset in [11]) 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 [3], 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 very significant impact this sampling bias can cause, we set up the following experiment (more details and variations in Section 2).

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



(a) Pansiot-Grad (b) Induced Subgraph on $G_{N,p}$

Figure 1: Power Law 2: Frequency Vs Degree

is beyond the scope of this experiment; we simply assign edges random weights $1 + \epsilon$, where ϵ is chosen uniformly at random from $[-\frac{1}{|V|}, \frac{1}{|V|}]$ and use shortest-path routing (the random weights are chosen solely to break ties between shortest-path routes).

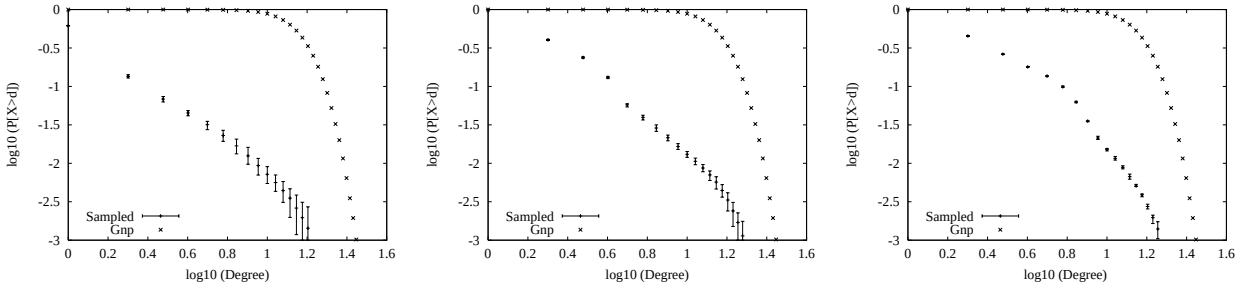
In Figure 1(b), we present a frequency vs. degree plot on log-log scale of the induced subgraph for $k = 1, m = 1000, N = 100,000, Np = 15$ (where Np is the average degree of a vertex). These parameters were chosen specifically to provide visual similarity to the plot from the dataset in [11]; we report on similar results for many other parameter settings later in the paper. While the induced subgraph demonstrates an equally striking frequency vs. degree power-law fit, this is a *measurement artifact* and is *not* a reflection of the underlying random graph. As is shown in the lower plot of Figure 1(b) (now depicted as a ccdf on log-log axes), the degree distribution of the underlying random graph is far from a power-law (it is well-known to be Poisson), while the degree distribution of the sampled graph exhibits a fit surprisingly like a power-law.

These plots 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? Are there methods which can reveal the presence of sampling bias in a traceroute dataset?

We explore the sources and effects of sampling bias in several stages. First, in Section 2, we conduct a thorough investigation of sampled subgraphs on *generated topologies*, namely classical random graphs and power-law random graphs (PLRGs), that expands upon and develops the arguments presented earlier in the introduction. We then explore the nature of graph sampling bias analytically in Section 3 and formulate tests to detect the presence of sampling bias. Then, in Section 4, we consider traceroute-based mapping studies in the Internet, and consider the evidence for

and against the sampling bias effects observed in Section 4.

2 Examining Node Degree Distribution of Sampled Subgraphs



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

Figure 2: Degree Distribution of Subgraph sampled from underlying $G_{N,p}$ ($N = 100,000$, $p = 0.00015$)

As briefly introduced earlier, sampled subgraphs can exhibit degree distributions that can deviate 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 [5] and power-law random graphs derived from the PLRG model [1]. We then examine possible sources of the bias responsible for highly variable degree distributions in sampled subgraphs.

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

2.1 Definitions and Assumptions

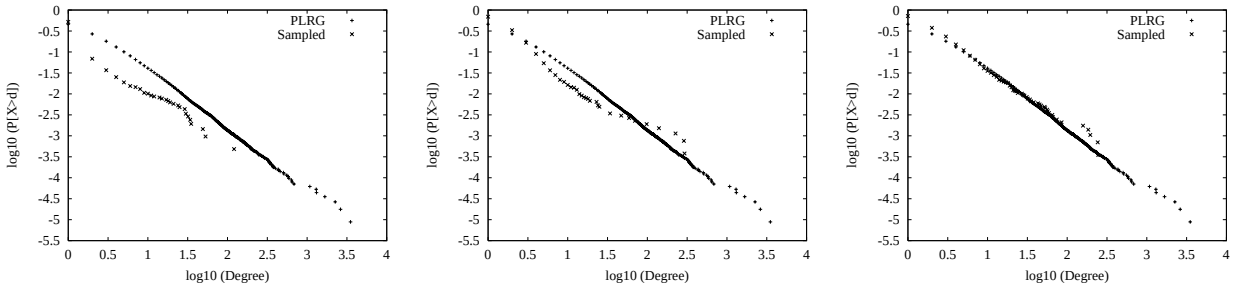
Let $G = (V, E)$ be a given sparse undirected graph with $|V| = N$. Our experimental methodology assigns random real-valued weights to the edges as follows: for all edges e in E, let the link weight $w(e) = 1 + \epsilon_e$ where ϵ_e is chosen uniformly and independently for each edge from the interval $[-\frac{1}{N}, \frac{1}{N}]$ (The noise attributed to edge weights is used solely to break ties in subsequent shortest path computation.) Then assume that we have k distinct source vertices selected at random, and m distinct destination vertices also selected at random. For each source-destination pair, we compute the shortest path between the source and destination. Then, 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.

As discussed in the Introduction, this experimental setup is motivated by traceroute-driven studies for mapping and understanding the Internet topology. In those studies, point-to-point measurements conducted to a large set of destinations from a set of distributed vantage points 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, etc. But it does model a crucial point: the fact that such a collection of traceroutes samples nodes and edges within the underlying topology *unevenly*.

2.2 Evidence for Power-Laws in Sampled Random Graphs

We begin by presenting experimental evidence, using two choices of underlying graphs: graphs generated by the classical Erdös-Rényi random graph model [5] and graphs generated by the power-law random graph (PLRG) model [1]. As we describe in more detail momentarily, 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 (as the name implies), the degree distribution of PLRG graphs follows a power-law.

Our first set of experiments employ the classical random graph model for the the underlying G. Briefly, this means that G is a graph with N labelled nodes such that every distinct pair of nodes is directly connected with probability p. A graph constructed in this manner is traditionally denoted by $G_{N,p}$. In all the random graphs we consider, Np is sufficiently large that the graph is connected with high probability. For our experiments, we ensured that each graph generated was connected.



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

Figure 3: Degree Distribution of Subgraph sampled from underlying PLRG Graph

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 except Figure 2(c).[1]

The results presented in these plots are important for three 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 such, conclusions that are made about $\hat{G}$ (*e.g.*, explanations put forward to explain the measured degree distribution [6]) may not necessarily apply to the underlying G. Third, even with a relatively large number of sources we cannot capture the true degree distribution, which highlights the inherent inefficiencies of this style of topology measurement.

2.3 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 [1] (PLRG).

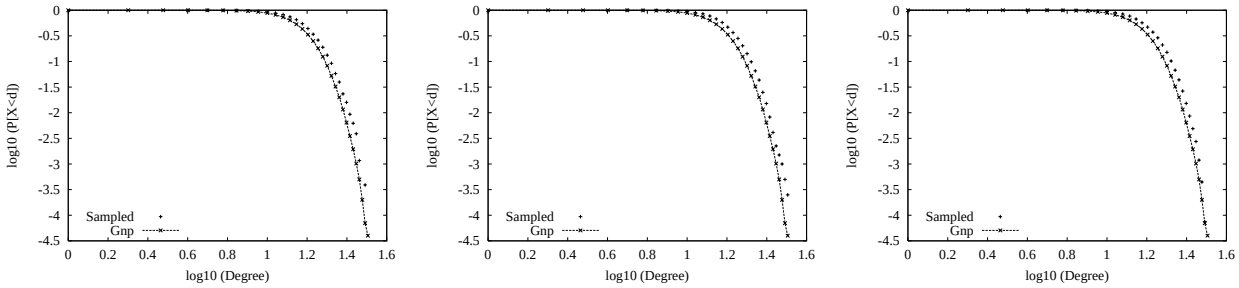
[1]Figure 2(c) shows 80% intervals for 10 trials. Due to time constraints, we could not complete 100 trials in this case.

Briefly, given N nodes and exponent β, the PLRG model initially assigns degrees drawn from a power-law distribution with exponent β and then proceeds to interconnect the nodes as follows. For each node n with target degree d, n is cloned d times and the resulting nodes are connected via a random matching. Cloned vertices are then collapsed. After removing self-loops and multi-edges and extracting the the largest connected component, our underlying power-law graph has $112,959$ nodes and $186,629$ edges with power-law exponent of about 2.1 (to match the exponent discovered by [7] in their router dataset). Figure 3 shows the degree distribution of $\hat{G}$ induced by 1, 5, 10 sources to 1000 destinations. Here, the sampled graph $\hat{G}$ exhibits a degree distribution visually similar to the underlying G. This is clearly in contrast to our earlier experiments on $G_{N,p}$, where we found that the sampled $\hat{G}$ exhibited a statistically distinct distribution. Further, even one source is sufficient to produce a degree distribution similar to that of the underlying PLRG graph.

2.4 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 what produces this biases for $G_{N,p}$ graphs and by doing so, provide some reasons for the emergence of long-tailed degree distribution in sampled graphs. Our explanations stem from observations of extensive simulations; we subsequently present an analytical justification.

An initial hypothesis to explain this phenomena is that the shortest path routing algorithm favors the higher degree nodes of $G_{N,p}$ in the computed optimal paths. In such a scenario, the high degree nodes of $G_{N,p}$ reduce the distance to reach destination nodes and so become frequently explored intermediate nodes. To test this hypothesis, 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 $G_{N,p}$.

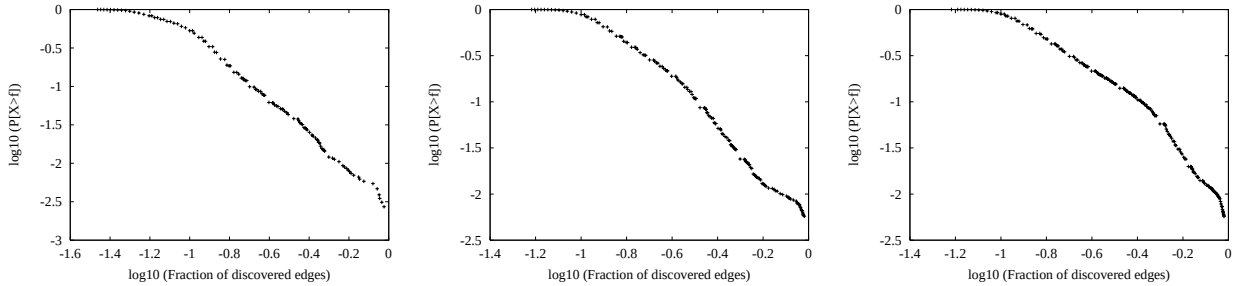


(a) 1 source, 1000 destinations (b) 5 sources, 1000 destinations (c) 10 sources, 1000 destinations

Figure 4: Hypothesis 1: Does shortest path routing select high degree nodes in $G_{N,p}$ graphs?

Figure 4 plots this 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 second hypothesis for the source of sampling bias concerns edges. One consequence of taking measurements using only a small number of sources is that edges are disproportionately selected. Therefore, another possible source of bias is in the omission of edges incident to a node in $\hat{G}$. To test this hypothesis, we examine the fraction of edges in our underlying $G_{N,p}$ that are



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

Figure 5: Hypothesis 2: Does edge omission contribute to a long tailed degree distribution in $G_{N,p}$ graphs?

discovered incident to each node n in $\hat{G}$. This is shown in Figure 5, where fraction of a node's true edges that are observed is plotted as a CCDF on log-log axes. If all edges incident to a node n of $\hat{G}$ were discovered, we would expect to see a horizontal line at probability 1. Instead, Figure 5 shows that most of the nodes have a very small fraction of edges discovered and only a few nodes (the high degree nodes of $\hat{G}$) have all their edges discovered. These plots support our second hypothesis: the skewed degree distribution of $\hat{G}$ is an artifact arising because of omitted edges.

3 Analysis and Inference

In this section 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.

3.1 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 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 2.1. 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:*

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

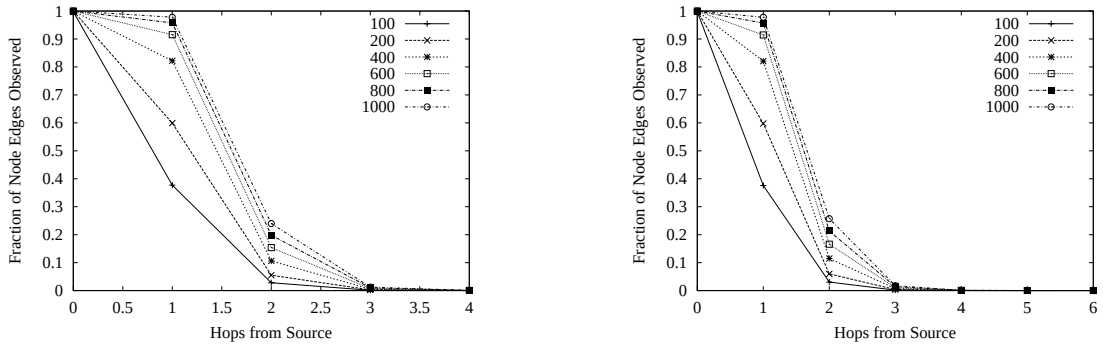


Figure 6: Visibility of Edges with Varying Number of Destinations; (a) N = 10,000; (b) N = 1,000,000.

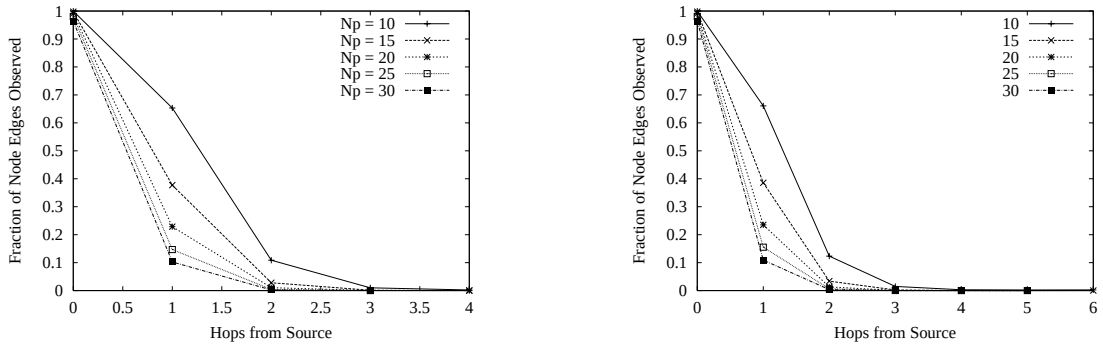


Figure 7: Visibility of Edges with Varying Vertex Degree in G; (a) N = 10,000; (b) N = 1,000,000.

and

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

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.

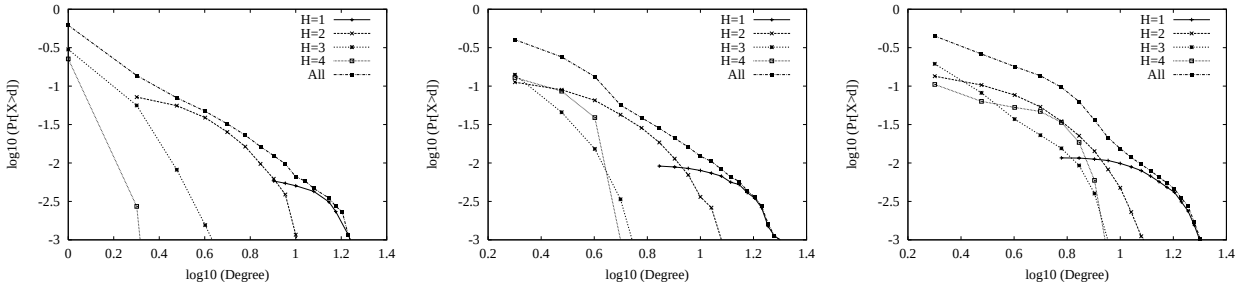
For the proof of Theorem 1 see Appendix A.

In order to evaluate this expression we need $|\Gamma_h|$. In [4], a number of bounds are given for $|\Gamma_h|$, and similar results are developed in [16]; however in general, tight bounds for this expression over the entire graph are not known. As a result we use an approximation to $|\Gamma_h|$ derived from simulation and consistent with the bounds derived in [4, 16].

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

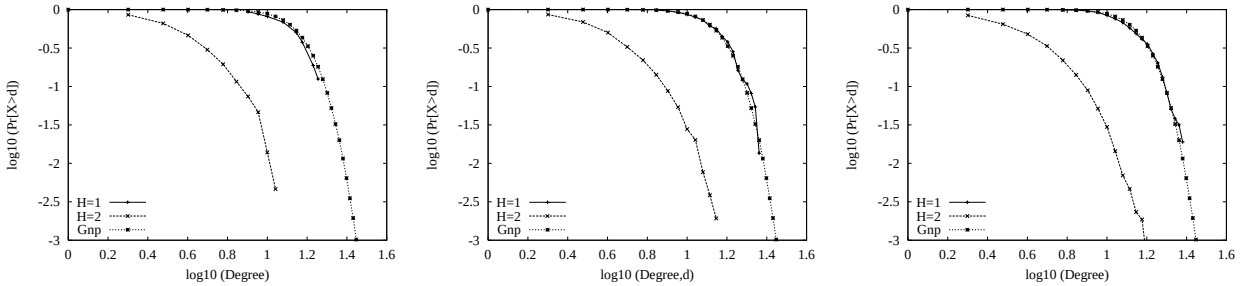
from the source. When this probability is small, we are missing most edges and so our estimates of node degree will be very inaccurate.

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.



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

Figure 8: LLCD of $Pr[D|H] * Pr[H]$ for sampled subgraphs of $G_{N,p}$



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

Figure 9: LLCD of $Pr[D|H]$ for hops 1 and 2 with true distribution

The plots show that over the vast majority of nodes in $\hat{G}$, visibility of edges is abysmal. 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 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.

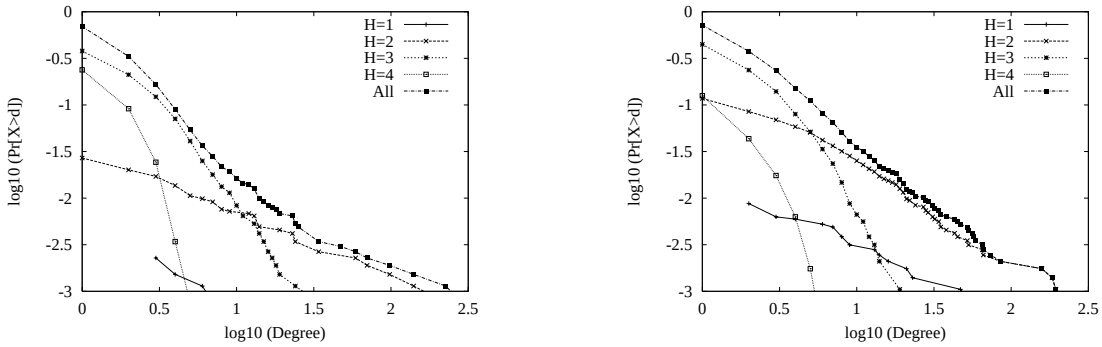
These results show that shortest-path trees only effectively explore a very small neighborhood around the source in a random graph. This helps explain the effect observed in Figure 5. Further-

more, 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 more formally and use it to examine graphs derived from traceroute measurements.

3.2 Inferring the Presence of Bias

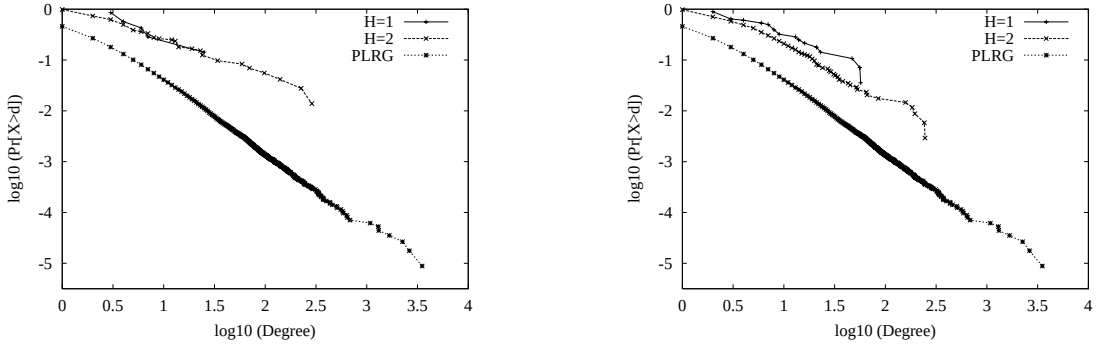
In the previous subsections, we provided evidence for and identified sources of bias when sampling G via shortest-path trees. Given these findings, a natural question to ask is if it is possible to detect evidence of bias in similar measurements when the underlying topology is unknown.

We start from the observation made in the last subsection, which showed that nodes close to the measurement source were explored much more thoroughly 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. However, if measurements are biased, we should be able to detect that by looking at statistics as a function of distance from source.



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

Figure 10: LLCD of $Pr[D|H] * Pr[H]$ for 5 and 10 sources for PLRG



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

Figure 11: LLCD of $Pr[D|H]$ for 5 and 10 sources for Hops 1 and 2 with true distribution

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 $Pr[D|H]$.

We first study $Pr[D|H]$ when $\hat{G}$ is sampled from $G_{N,p}$. Figure 8 shows how node degrees at each hop conspire to produce the illusion of an overall power-law degree distribution of $\hat{G}$. We make two observations. First, the highest degree nodes are found at $H = 1$, that is, at hops nearest to the source nodes. Second, $Pr[D|H = 1]$ appears to be visually short-tailed and is very different from both the overall degree distribution and the degree distributions at larger hop distances. These two observations suggest criteria for detecting bias in topology measurements of an unknown graph. These criteria are:

C1 *Are the highest-degree nodes 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.

A corollary of these observations is that in the presence of bias, $Pr[D|H = 1]$ should best approximate the true degree of the underlying G. This is verified in Figure 9 which shows that in the $G_{N,p}$ case, the distributional shape for hop 1 nodes is nearly indistinguishable from the true distribution — while for hop 2 nodes the difference is sharp.

If these criteria are to be useful they should hold for the case of power-law random graphs as well. Figure 10 shows $Pr[D|H]$ for graphs sampled from the PLRG generated 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 $H = 3$. Finally, Figure 11 shows that just as in the case for $G_{N,p}$ graphs, the degree distributions of the nodes at hops nearest to the source visually approximate the underlying degree distribution well.

The consistent behavior of our 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. We emphasize that meeting either or both criteria does not conclusively demonstrate sample bias, but rather is consistent with the existence of sample bias. On the other hand, a dataset meeting both criteria would seem to be a poor choice for use in making generalizations about the true nature of the underlying graph.

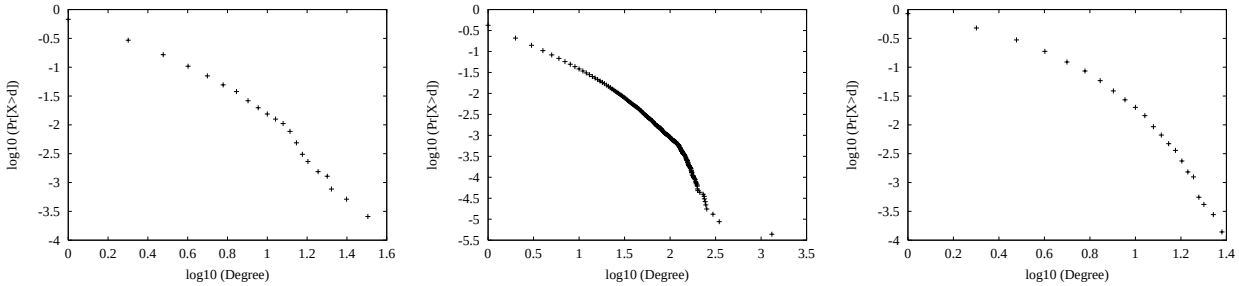
4 Examining Node Degree Distribution of Traceroute Datasets

To gauge the extent of potential sample bias in measurements of router-level graphs, we now turn to examining existing IP topology measurements.

4.1 From Models to Datasets

Before turning to empirical data, it is helpful to assess the ways in which real data differs from the experiments we have described so far.

An example of the state of the art in topology measurement is CAIDA's Skitter project [8], which consists of 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:



(a) Pansiot-Grad Routers (b) Mercator Routers (c) Skitter Routers

Figure 12: Degree distribution of datasets

1. We have assumed that sources and destinations are randomly placed in the graph. In a real measurement system, destinations may not be truly randomly chosen, although their locations are generally not tightly constrained. However the location of sources in particular is constrained by the mechanics of setting up active measurement sites. What effect can this have on our results? It may be that the neighborhoods around sources are unusual for this reason; while this is a new and different form of bias it would nevertheless be detected by our criteria.

2. 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 [17, 18], it does not reflect the inflating effect that routing policy has on paths in the Internet [15, 14]. However routing policy is designed in general to find *short* paths (typical paths in the Internet are on the order of 12-18 hops), and the kinds of sampling bias we consider here would seem to be present in any system trying to keep paths short.

4.2 Datasets

We use three different snapshots of the router topology collected at different time periods.[2] Table 1 summarizes these datasets and Figure 12 reproduces their node degree distribution as a log-log CCDF. Our first dataset, *Pansiot-Grad routers*, dates from 1995 [11]. It was this dataset that was first used as evidence for power-law router degree distribution in the paper by Faloutsos *et al* [7]. A subsequent and much larger dataset, *Mercator Routers*, was collected in August 1999 and appeared in [9]. The authors of [9] also found evidence for a power-law degree distribution.[3] Our third dataset, obtained from 8 distinct sources of the Skitter project after resolving interfaces to routers (examined in [2]) also shows evidence for a long-tailed degree distribution.

Our measured datasets are significantly larger than our sampled graphs from simulations and so have much longer typical path lengths. As a result the number of nodes at a given hop distance from the source in the measured datasets is a much smaller fraction of the total node set than in

[2]Perhaps the largest IP topology snapshots are recent measurements from the Skitter system, *e.g.*, [3]. Unfortunately these datasets do not resolve interfaces to routers, and so introduce another serious and complicating source of bias in trying to assess router degree distribution.

[3]To be precise, the authors of [9] 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.

the simulation graphs. Thus, for the two larger datasets, it is necessary to aggregate nodes across hops so as to create "rings" $h_i \leq h < h_j$ containing enough nodes to form a smooth empirical distribution function

For the *Mercator Routers* dataset, 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 information is not available for this dataset. For all other datasets, we have IP path information and rely on it to compute hop distance. For all datasets, H denotes the minimum hops to a node from the source nodes.

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 1: Summary of Datasets Examined

4.3 Detecting Bias

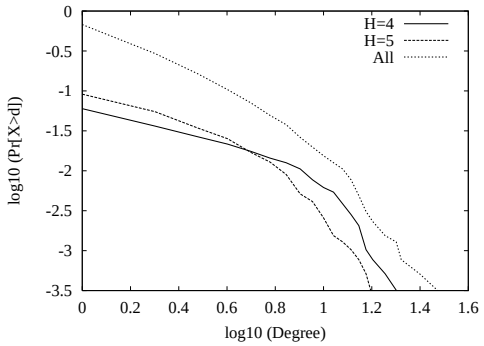
We now proceed to apply our criteria for sample bias to each of these datasets.

Pansiot-Grad Routers. **C1:** In examining the Pansiot-Grad dataset, we find that the highest degree routers were indeed close to the source. **C2:** Furthermore, the degree distribution for routers at different distances is shown in Figure 13(a). This figure shows evidence that the degree distribution closer the source $H = 4$ is different from the degree distribution farther from the source $H = 5$. We conclude that this dataset shows evidence consistent with sampling bias, and hence that it may not accurately represent the true degree distribution of the underlying graph.

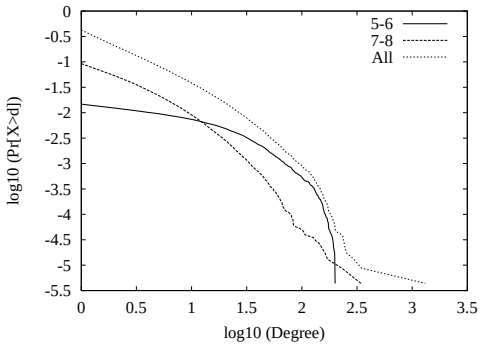
Mercator Routers. **C1:** Likewise, for the Mercator dataset, the highest degree routers are again found close to the source. **C2:** In examining the conditional degree distributions (Figure 13(b)) we see strong evidence of a difference in distribution as a function of distance from the source. We conclude that this dataset also shows evidence consistent with sampling bias.

Skitter Routers. **C1:** In the Skitter dataset, we find that the higher degree nodes appear later in the traceroute paths, at intermediate hops. **C2:** The conditional degree distributions for the Skitter dataset is shown in Figure 13(c). We conclude that the evidence from this figure is indeterminate. Thus we conclude that evidence for bias in the Skitter dataset is not as strong as in the other two.

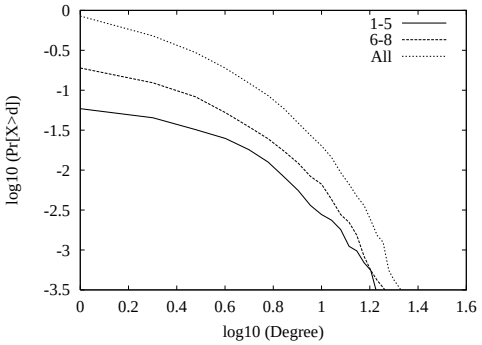
We summarize the results from the analysis in this section in Table 2. Overall we find that the first two datasets, and perhaps the third, do not provide strong bases for conclusions about properties such as the degree distribution of the underlying graph.



(a) Pansiot-Grad Routers



(b) Mercator Routers



(c) Skitter Routers

Figure 13: Examining $Pr[D|H]$ for empirical datasets.

Dataset Name	**C1**	**C2**	Conclusion
Pansiot-Grad	Yes	Yes	Consistent with Bias
Mercator Routers	Yes	Yes	Consistent with Bias
Skitter Routers	No	?	Inconclusive

Table 2: Summary of Evidence for Bias in Empirical Data

5 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 those methods to various empirically captured router inventories suggests that evidence of sampling bias may well be present in a number of cases. Two out of the three datasets we

examined showed properties consistent with sampling bias, while the other dataset's properties are inconclusive.

Our results suggest that since long-tailed degree distributions can arise simply through biased sampling of graphs, and since capturing the true degree distribution is a difficult task without deploying a substantial number of sources, node degree distribution may not be sufficiently robust for characterizing [7] or comparing router-level topologies [10, 13].

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. For example, a sampling method with the capability to accurately sample the degree of a randomly chosen router in the Internet could be used as an alternative to mapping heuristics, and could help shed light on the true degree distribution of the underlying network.

References

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

[2] 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 (IMW '01)*, November 2001.

[3] A. Broido and K. Claffy. Connectivity of IP Graphs. In *Proceedings of SPIE ITCom '01, Scalability and Traffic Control in IP Networks*, August 2001.

[4] Fan Chung and Linyuan Lu. The diameter of random sparse graphs. *Advances in Applied Math*, pages 257–279, 2001.

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

[6] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically Optimized Tradeoffs. http://www.cs.berkeley.edu/christos/.

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

[8] Cooperative Association for Internet Data Analysis (CAIDA). The Skitter project. At *http://www.caida.org/Tools/Skitter*.

[9] R. Govindan and H. Tangmunarunkit. Heuristics for Internet Map Discovery. In *Proceedings of IEEE/INFOCOM'00*, March 2000.

[10] D. Magoni and J. Pansiot. Comparative Study of Internet-like Topology Generators. Technical Report Research Report ULP/LSIIT-RR-2001/08, Université Louis Pasteur, 2001.

[11] J. Pansiot and D. Grad. On Routes and Multicast Trees in the Internet. *ACM Computer Communication Review*, 28(1):41–50, January 1998.

[12] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP Topologies with Rocketfuel. In *Proceedings of ACM/SIGCOMM '02 (to appear)*, August 2002.

[13] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. Network Topology Generators: Degree-Based vs. Structural. In *Proceedings of ACM/SIGCOMM'02 (To appear)*, Pittsburgh, PA, August 2002.

[14] H. Tangmunarunkit, R. Govindan, and S. Shenker. Internet Path Inflation Due to Policy Routing. In *SPIE International Conference on Performance and Control of Network Systems*, 2001.

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

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

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

[18] P. van Mieghem and M. Janic. Stability of a Multicast Tree. In *Proceedings of IEEE INFOCOM'02*, 2002.

A Analysis of Shortest-Path Trees on Random Graphs

There is a single source and there are D destination nodes, all chosen randomly from the underlying graph which is a classical random graph $G = (V, E)$ with $|V| = N$ vertices. The probability that any two vertices form an edge is p. Edges in the graph have constant weight. We assume that $Np \gg \ln N/N$ so that the graph is connected.

We are interested in this problem: what is the degree distribution of nodes in the subgraph of G defined by the shortest-path tree from the single source to the n destinations? Our interest is in the case in which this subgraph is very small compared to G. When a destination has multiple equal-cost paths to the source, a path is chosen at random (however, we argue below that such events are rare).

We take a fairly direct, counting or combinatorial approach to the solution. Throughout we will denote the binomial distribution with success probability p over n trials as $B(n, p)$ and this distribution's value at x as $B(n, p, x)$.

A.1 Probability of traversing an edge at distance h

As a first step, we are interested in the probability that the shortest path to n destinations ($n \leq D$) passes through a given edge of a given node at h hops from the source. We denote this as $p_h(n)$.

A.1.1 Base Case

Consider the source node (the root of the tree).

First of all, some of the destinations may be the same as the root. The probability that a single destination is the same as the root is $1/N$ and so the number of destinations that are the same as the root is a random variable I with distribution $B(D, 1/N)$.

For any destination node that is not the same as the root, its shortest path to the root is equally likely to pass through any of the edges incident on the root (recall that all edges have equal weight). So if the root node happens to have A edges and there are M destinations that are not the same as the root we can write

$$p_0(n \mid A \text{ edges, } M \text{ destinations}) = B(M, 1/A, n) = C_n^M (1/A)^n (1-(1/A))^{M-n} \quad \text{for } n = 0, ..., M.$$

However the number of destinations not yet visited is a random variable; it is $D - I$. Taking this into account we have:

$$p_0(n \mid A \text{ edges}) = \sum_{i=0}^{D} B(D, 1/N, i) B(D - i, 1/A, n) \quad \text{for } n = 0, ..., D.$$

and finally for an arbitrarily chosen root node we have

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

where $P(\lambda, j)$ is the Poisson distribution, used here to describe the probability of a node having j edges in a random graph with average degree λ.

A.1.2 Recursive Step

Now consider an arbitrary node at h hops away from the root. The number of destinations whose shortest paths pass through this node is a random variable with distribution $p_{h-1}(n)$. For any destination that is not the same as the given node, we take the approximation that its shortest path to the root is again equally likely to pass through any of the node's edges (except for the path to the root).

This approximation is justified on the basis that the shortest-path subtree being constructed is negligibly small compared to the graph itself. So, while some of the node's edges may connect to other nodes in the shortest-path tree, causing a dependence among nodes that is not captured by this assumption, we take the probability of that event to be very small and so ignore it.

First we count the number of destinations that may be identical to the given node. Let us denote the set of nodes that are distance k from an arbitrary node as $\Gamma_k(x)$; that is,

$$\Gamma_k(x) = \{y \in G \mid \textit{distance}(x, y) = k\}.$$

Then $|\Gamma_h(x)|$ is the number of nodes at hop h from the source. The probability that one destination is identical to the given node is then $|\Gamma_h(x)|/N$ and the probability that m nodes out of n are identical to the given node is $B(n, |\Gamma_h(x)|/N, m)$

So we can write

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

In particular, the probability that an edge that is incident on a node in the tree at depth h is actually part of the tree is $1 - p_h(0)$.

A.2 Degree Distribution of Subgraph

Now, let us denote the probability that a node at hop h has degree i in the subgraph as $d_h(i)$.

If the node has A edges in G, this probability is

$$d_h(i \mid A \text{ edges}) = B(A, 1 - p_h(0), i)$$

and the probability for an arbitrary node at hop h is

$$d_h(i) = \sum_{j=0}^{\infty} P(Np, j) B(j, 1 - p_h(0), i).$$

Therefore the overall degree distribution is:

$$P[\text{node in subgraph has degree } i] = \sum_{h=0}^{\text{diameter}} p(h) d_h(i)$$

where $p(h)$ is the probability of finding a node in G at distance h from the source in the subtree.

The average degree of a node at hop h in the subtree is

$$E_h[d] = \sum_{i=0}^{\infty} i \cdot d_h(i).$$

So we form a crude estimate of $p(h)$ (a better one is possible but expensive) as

$$p(h) \approx \frac{\Pi_{j=1}^{h} E_j[d]}{\sum_{k=1}^{\infty} \Pi_{j=1}^{k} E_j[d]}$$