Bounding the Bias of Tree-Like Sampling in IP Topologies

Reuven Cohen [*] Mira Gonen[†] Avishai Wool[‡]

December 9, 2006

Abstract

It is widely believed that the Internet's AS-graph degree distribution obeys a power-law form. Most of the evidence showing the power-law distribution is based on BGP data. However, it was recently argued that since BGP collects data in a tree-like fashion, it only produces a sample of the degree distribution, and this sample may be biased. This argument was backed by simulation data and mathematical analysis, which demonstrated that under certain conditions a tree sampling procedure can produce an artificial power-law in the degree distribution. Thus, although the *observed* degree distribution of the AS-graph follows a power-law, this phenomenon may be an *artifact* of the sampling process.

In this work we provide some evidence to the contrary. We show, by analysis and simulation, that when the underlying graph degree distribution obeys a power-law with an exponent $\gamma > 2$, a tree-like sampling process produces a negligible bias in the sampled degree distribution. Furthermore, recent data collected from the DIMES project, which is not based on BGP sampling, indicates that the underlying AS-graph indeed obeys a power-law degree distribution with an exponent $\gamma > 2$. By combining this empirical data with our analysis, we conclude that the bias in the degree distribution calculated from BGP data is negligible.

Classification: Network Computing, Internet Topology Models

1 Introduction

1.1 Background and Motivation

The connectivity of the Internet crucially depends on the relationships between thousands of Autonomous Systems (ASes) that exchange routing information using the Border Gateway Protocol (BGP). These relationships can be modeled as a graph, called the AS-graph, in which the vertices model the ASes, and the edges model the peering arrangements between the ASes.

Significant progress has been made in the study of the AS-graph's topology over the last few years. A great deal of effort has been spent measuring topological features of the Internet. Numerous research projects [FFF99, AB00, CCG$^+$02, LBCX03, WGJ$^+$02, WJ02, BT02, BGW05, BGW04, KRR01, SS05, LC03, BB01, BBBC01, RN04, TPSF01, TGJ$^+$02, GT00, BS02] have ventured to capture the Internet's

[*] Reuven Cohen is with the Deparment of Electrical and Computer Engineering, Boston University, Boston, MA 02215, USA. cohenr@shoshi.ph.biu.ac.il

[†] Mira Gonen is with the School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978, Israel. gonenmir@tau.ac.il

[‡] Avishai Wool is with the School of Electrical Engineering , Tel Aviv University, Ramat Aviv 69978, Israel. yash@acm.org

topology. Based on these and other topological studies, it is widely believed that the Internet's degree distribution has a power-law form with an exponent $2 < \gamma < 3$, i.e., the fraction of vertices with degree k is proportional to $k^{-\gamma}$. Most of these studies are based on BGP data from sources such as RouteViews [Rou05] and CIDR Reports [BSH03].

1.2 Related Work

It was recently argued [LBCX03, CM04a, CM04b, PR04, ACKM05, SS05, CCG$^+$02, TGJ$^+$02, WGJ$^+$02] that the evidence obtained from the analysis of the Internet graphs sampled as described above may be biased. In a thought-provoking article, Lakhina et al. [LBCX03] claimed that a power-law degree distribution may be an artifact of the BGP data collection procedure. They suggest that although the *observed* degree distribution of the AS-graph follows a power-law distribution, the degree distribution of the real AS-graph might be completely different. They claim that with tree-like sampling, such as that employed by BGP, an edge is much more likely to be visible, i.e., included in the sampled graph, if it is close to the root. Moreover, in tree-like sampling, high-degree vertices are more likely to be encountered early on, and therefore they are sampled more accurately than low-degree vertices. They backed this argument with simulations that indicated that under some conditions, a BFS sampling process *in itself* is sufficient to produce a power-law degree distribution in the sample, even when the underlying graph is a sufficiently-dense Erdős-Renyi [ER60] graph.

Subsequently [CM04a, CM04b] gave a mathematical foundation to the argument of [LBCX03]. They showed that BFS tree sampling produces a power-law degree distribution, with an exponent of $\gamma = 1$, both for Poisson-distributed random graphs and for δ-regular random graphs. In other words, they showed that a tree sampling process may have a significant bias, and may produce an artificial power-law—albeit with an exponent that is very different from that observed in the AS-graph.

On the other hand, Petermann and De Los Rios [PR04] showed that for single-source tree sampling of a BA graph [BA99], the exponent obtained for the power-law distribution is only slightly under-estimated. This cannot be viewed as strong evidence against the argument of [LBCX03], since the analysis assumes a BA-model, which is a highly idealized evolution model of the AS-graph. However, this result does indicate that at least in some power-law graphs, a tree sample does not create a significant bias in the degree distribution.

More recently, [ACKM05] analyzed the degree distribution discovered by a BFS tree sampling process over a general graph. Among other results, they gave a general, but rather unwieldy, expression of the degree distribution of the sampled graph, depending on the underlying graph degree distribution. This work is the starting point of our analysis: we use the results of [ACKM05] to analyze the sampled degree distribution when the underlying graph has a power-law degree distribution.

A new development in the empirical measurement of the Internet topology was suggested recently by Shavitt and Shir [SS05]. In this work they describe an Internet mapping system called DIMES. DIMES is a distributed measurement infrastructure for the Internet that is based on the deployment of thousands of light weight measurement agents around the globe. Unlike BGP data, that is sampled in a tree-like fashion, DIMES executes `traceroute`s among all pairs of its agents, collects the router-level results, and aggregates the AS-graph from this data. Because of this measurement methodology, DIMES discovers significantly more links than BGP-based systems. The salient point for our purposes is that DIMES data too shows a power-law degree distribution, with an exponent $2 < \gamma < 3$ — and since the DIMES system uses an all-pairs measurement paradigm, it is difficult to claim that the power-law is an artifact of a tree-like sampling.

1.3 Contributions

Our main contribution is our analysis of the degree distribution observed in the BFS tree sample, when the underlying graph has a power-law distribution with an exponent $2 < \gamma < 3$. Under these conditions, we prove that the bias in the power-law is negligible: with high probability, the degree distribution of the high-degree nodes in the sample also exhibits a power-law, with exactly the same exponent γ. We validate our mathematical analysis with simulation results using the DIMES-measured AS-graph as the underlying power-law graph.

Putting this result in the context of the Internet topology, we recall the data collected from the DIMES project is not based on BGP-style tree sampling. Nevertheless, DIMES data indicates that the underlying AS-graph indeed obeys a power-law degree distribution with an exponent $\gamma > 2$. By combining this empirical data with our analysis, we conclude that the bias in the degree distribution calculated from BGP data is negligible.

Organization: In the next section we give an overview of the results of [ACKM05] we rely on. In Section 3 we show our main result, that the bias in the degree distribution of a tree-sampled power-law graph is negligible. In Section 4 we sketch an alternative, more rigorous, analysis of a weaker result, that validates some of the approximations we used in our main result. Section 5 describes the results of our simulations. We conclude with Section 6.

2 Highlights of [ACKM05]

2.1 The General Framework

The proof of our result is based on the model, sampling process, and the main results described in [ACKM05]. In this section we give a brief introduction to the main results we need.

Notation Throughout the paper we use $G = (V, E)$ to denote the underlying graph, and $n = |V|$ to denote the number vertices.

Definition 1 *We say that* $\{a_j\}$ *is a degree distribution of* G *if* G *contains* $a_j \cdot n$ *nodes of degree* j*.*

In the [ACKM05] model, the graph G is not a given graph but a random graph chosen out of a family of graphs obeying a given degree distribution $\{a_j\}$. The basic setting is the configuration model of [Bol85]: for each vertex of degree k we create k copies, and then define the edges of the graph according to a uniformly random matching on these copies.

2.2 The BFS Tree Sampling Process

The [ACKM05] model defines a randomized process, that simultaneously produces a random graph G obeying the degree distribution $\{a_j\}$, and a BFS tree T that represents the sample. Note that for a given graph G, a BFS algorithm is a deterministic algorithm, but different outcomes are possible, depending on the order in which outgoing edges are traversed. In [ACKM05] model, a random choice determines this order.

The sampling process is thought of taking place in continuous time. However, for technical reasons the authors define a non-standard notion of time, which we denote by a capitalized word (Time). In this model, the BFS sample process starts at Time $t = 1$ with an empty tree T. As the sampling process evolves, Time decreases to $t = 0$, when the sample tree T includes all n nodes (assuming G is connected).

Before the process starts, for each vertex v there are $deg(v)$ *copies* of v. Each copy is given a real-valued *index* chosen uniformly at random from the unit interval $[0,1]$. Namely, vertex v has $deg(v)$ indices, chosen uniformly and independently at random from the unit interval $[0,1]$. At every Time step t two copies are matched: one copy is a copy of a vertex already discovered, and the other copy is a copy with index t. Such a matched pair forms an edge of the original graph. According to [ACKM05] at Time t the indices of the unmatched copies are uniformly random in $[0,t)$. Let the *maximum index* of a vertex be the maximum of all its copies' indices. Then at any Time t, the vertices that have not been discovered yet are precisely those whose maximum index is less than t. An edge will be *visible*, namely, included in the BFS tree, if at the Time its endpoints are matched one of them is a copy of an undiscovered vertex.

2.3 Useful Notations and Theorems of [ACKM05]

Let v_t be the vertex that has a copy with maximum index t. Denote by $P_{vis}(t)$ the probability that *another* edge outgoing from v_t appears in the BFS tree. Namely, v_t is the vertex that was discovered at Time t, and we are interested in the probability that *another* vertex is discovered through v_t.

Using the expectation of $P_{vis}(t)$ as approximation we get from Equation (7) and Lemma 3 of [ACKM05] the following Theorem: [1]

Theorem 1 *Let G be a connected graph and let $\{a_j\}$ be a degree distribution that is upper bounded by a power-law with an exponent larger than 2. Let $\mu = \sum_j ja_j$ be the mean degree of G. Then it holds that*

$$P_{vis}(t) \quad \approx \quad \frac{1}{\sum_j ja_jt^j}\sum_k ka_kt^k\left(\frac{\sum_j ja_jt^j}{\mu t^2}\right)^k. \qquad (1)$$

3 The Degree Distribution of the Sampled Graph

Our goal in this section is to show that the bias observed in BFS tree sampling regarding the degree distribution of the sampled graph is not significant, when the underlying graph degree distribution obeys a power-law with an exponent $\gamma > 2$. We show the above by examining the BFS tree received by the sampling process of [ACKM05], described in the previous section. Finding a BFS tree from a single source is an idealization of the BGP data collection process. Thus, if the bias is negligible when using such BFS process, then we argue that it is very likely to be negligible when using the more general case of BGP.

Recall that we focus on a BFS process on a random graph G. Let T be the BFS tree received by this process. Let $deg_T(v)$ denote the degree of node v in the BFS tree received from the sampled graph. Let $deg_G(v)$ denote the degree of node v in the graph.

Definition 2 *We say that a vertex v has a* high-degree *if $deg_G(v) \geq 18$. We say that an event occurs* with high probability *(w.h.p) if it occurs with probability at least 0.16.*

The main theorem we prove is the following:

[1] In Lemma 3 of [ACKM05] there is a requirement that $a_j = 0$ for $j < 3$. This implies that with high probability the graph is connected. However, if we assume that the graph is connected, which is the case we are interested in, this requirement can be relaxed.

Theorem 2 *If the underlying graph degree distribution obeys a power-law with an exponent γ, where $2 < \gamma < 3$, then with w.h.p the degree distribution of the high-degree vertices of the sampled graph also follows a power-law, with the same exponent value.*

We prove Theorem 2 using several Lemmas and Theorems. Throughout the remainder of this paper we use the following setting:

Definition 3 *Let $a_k = C \cdot k^{-\gamma}$ be a degree distribution, where $2 < \gamma < 3$ and $C > 0$ is a constant normalization factor. Let μ denote the mean graph degree, i.e., $\mu = E[deg_G(v)] = \sum_j ja_j$.*

Note that since the degree distribution is a power-law with an exponent larger than 2, in this case μ is finite. Our starting point is Theorem 1 [ACKM05]. Our first step is to approximate the following sum, which appears in Equation (1), when the degree distribution obeys the power-law of Definition 3:

$$\sum_k ka_kt^k = C\sum_k kk^{-\gamma}t^k = C\sum_k k^{1-\gamma}t^k. \tag{2}$$

Lemma 1 *Let μ, a_k, and γ be as in Definition 3. Then*

$$\sum_k k^{1-\gamma}t^k \approx \frac{t^3}{\gamma-2},$$

for all $0 \leq t \leq 1$.

Proof:

$$\sum_k k^{1-\gamma}t^k \approx \int_1^\infty x^{1-\gamma}t^x dx$$

Let $g(t) = \int_1^\infty x^{1-\gamma}t^x dx$. Then for all $i \geq 1$ we have that the ith derivative of $g(t)$ is

$$g^{(i)}(t) = \int_1^\infty x^{1-\gamma}\prod_{j=0}^{i-1}(x-j)t^{x-i}dx.$$

Let us now evaluate g, and its derivatives, at the boundaries of $[0,1]$:

- $g(0) = 0$,
- $g^{(i)}(0) = 0$ for all $i \geq 1$,
- $g(1) = \int_1^\infty x^{1-\gamma}dx = \frac{1}{2-\gamma}x^{2-\gamma}\Big|_1^\infty = \frac{1}{\gamma-2}$,
- $g^{(i)}(1) = \infty$ for all $i \geq 2$.

We will approximate $g(t)$ by an interpolation polynomial $f(t) = \sum_\ell b_\ell t^\ell$. Obviously, no polynomial has $f^{(i)}(1) = \infty$ for any i. Thus we will use g's values at $t = 0, 1$, and the derivatives at 0. The minimal-degree non-trivial polynomial we can use is a cubic $f(t) = b_0 + b_1t + b_2t^2 + b_3t^3$. Since $g(0) = 0$ we get that $b_0 = 0$. $g'(0) = 0$ implies $b_1 = 0$, and $g''(0) = 0$ implies $b_2 = 0$. Since $g(1) = \frac{1}{\gamma-2}$ we get that $b_3 = \frac{1}{\gamma-2}$. Thus $f(t) = t^3/(\gamma-2)$ and

$$g(t) = \int_1^\infty x^{1-\gamma}t^x dx \approx \frac{t^3}{\gamma-2}. \quad \blacksquare \tag{3}$$

Notes:

- Lemma 1 approximates the sum of Equation (2) using a cubic polynomial. This is a somewhat arbitrary choice: one can use any polynomial of degree $d \geq 3$ and obtain similar results. Using higher-degree polynomials yields a better-quality approximation around $t = 0$ since additional derivatives are approximated, but does not necessarily improve the accuracy of the approximation at $t = 1$ since we can only use $g(1)$ itself. Since we mostly care about the early stages in the evolution, near Time $t = 1$, we only present the result for the special case of $d = 3$.

- At present we give no bound on the error of our polynomial approximation. Instead, to validate our results, in Section 4 we present a weaker result proven rigorously without using the approximation.

- No polynomial can to give $g^{(i)}(1) = \infty$ for *any* i, so our approximation accuracy is fundamentally limited around $t = 1$. It would be interesting, and technically more difficult, to approximate the sum using rational functions or other functions with an asymptote at $t = 1$. We leave this to future work.

Lemma 2 *Let μ, a_k, and γ be as in Definition 3. Then $\mu(\gamma - 2) \approx C$.*

Proof:

$$\begin{aligned}
\mu(\gamma-2) &= (\gamma-2)\sum_k k a_k = (\gamma-2)C\sum_k k k^{-\gamma} \\
&= (\gamma-2)C\sum_k k^{1-\gamma} \approx (\gamma-2)C\int_1^\infty x^{1-\gamma}dx \\
&= (\gamma-2)C\ \frac{1}{2-\gamma}x^{2-\gamma}\bigg|_1^\infty = (\gamma-2)C\frac{1}{\gamma-2} = C\ \ . \qquad (4)
\end{aligned}$$

(The last equality is valid for $\gamma > 2$). ■

In addition to being a building block in proving our main Theorem, the following Lemma 3 is important since it shows that most of the edges are detected early during in the sampling process, near Time $t = 1$.

Recall that $P_{vis}(t)$ is the probability that the vertex discovered at Time t gives rise to *another* edge in the BFS tree—i.e., not the edge it was detected with.

Lemma 3 $P_{vis}(t) \approx t^3$

Proof: Recall that $a_k = C \cdot k^{-\gamma}$. Using equation (1) and Lemma 1 we further approximate P_{vis}.

$$\begin{aligned}
P_{vis}(t) &\approx \frac{1}{\sum_j j a_j t^j}\sum_k k a_k t^k \left(\frac{\sum_j j a_j t^j}{\mu t^2}\right)^k \\
&\approx \frac{\gamma-2}{C\cdot t^3}\sum_k C\cdot k^{1-\gamma}\frac{1}{t^k \mu^k}\left(\frac{C\cdot t^3}{\gamma-2}\right)^k = \frac{\gamma-2}{t^3}\sum_k k^{1-\gamma}\left(C\cdot\frac{t^2}{\mu(\gamma-2)}\right)^k
\end{aligned}$$

and by Lemma 2 we have that

$$\approx \frac{\gamma-2}{t^3}\sum_k k^{1-\gamma}t^{2k}. \qquad (5)$$

Let $w = t^2$. By substituting w in equation (5) we get

$$P_{vis}(t) \approx \frac{\gamma-2}{t^3}\sum_k k^{1-\gamma}w^k \qquad (6)$$

Using Lemma 1 again we get that

$$P_{vis}(t) \quad \approx \quad \frac{\gamma-2}{t^3}\frac{w^3}{\gamma-2} = \left(\frac{w}{t}\right)^3 = t^3 \tag{7}$$

■

Note that at Time $t = 0$ Lemma 3 gives $P_{vis} \approx 0$, which is as expected: at the end of the BFS process no new tree-edges are detected. Furthermore, at Time $t = 1$ we get $P_{vis} \approx 1$, again matching our intuition that at the beginning of the BFS process the edges detected very likely to be tree edges. Moreover, observe that most edges are detected at the beginning of the BFS process.

Recall that in the BFS sampling process of [ACKM05] each copy of a vertex v is assigned a Time index $t \in [0, 1]$ (Section 2.2).

Definition 4 *Let* max-index(v) *be the maximum index of a vertex* v*, where the maximum is taken over all copies of* v*.*

Lemma 4 *Let* v *be a vertex with graph degree* i *and let* max-index$(v) = t$*. Then*

$$E[deg_T(v) \mid deg_G(v) = i, \text{max-index}(v) = t] \approx (i-1)t^3$$

Proof: We follow the discussion in [ACKM05], and neglect the possibility of self-loops and parallel edges involving a vertex v and its siblings, and ignore the fact that we are choosing without replacement (i.e., that processing each copy slightly changes the number of undetected vertices and the number of unmatched copies). Under these assumptions, the events that each of v's siblings give rise to edges that will be detected in the tree are independent, and [ACKM05] shows that the number of visible edges is approximately binomially distributed as $Bin(i-1, P_{vis}(t))$. Therefore

$$E\left[deg_T(v) \mid deg_G(v) = i, \text{max-index}(v) = t\right] = (i-1)P_{vis}(t). \tag{8}$$

Thus, using Lemma 3 it holds that

$$E\left[deg_T(v) \mid deg_G(v) = i, \text{max-index}(v) = t\right] \approx (i-1)t^3. \tag{9}$$

■

Theorem 3 *Let* v *be a vertex with graph degree* i*. Then*

$$E\left[deg_T(v) \mid deg_G(v) = i\right] \approx \frac{i(i-1)}{i+3} \tag{10}$$

Proof: Since $t = \text{max-index}(v)$ is the maximum of i independent uniform variables in $[0, 1]$, its probability density is $dt^i/dt = it^{i-1}$. Therefore, using Lemma 4, we get

$$\begin{aligned} E\left[deg_T(v) \mid deg_G(v) = i\right] &= \sum_k k\Pr\left[deg_T(v) = k \mid deg_G(v) = i\right] \\ &= \sum_k k \int_0^1 it^{i-1}\Pr\left[deg_T(v) = k \mid deg_G(v) = i, \text{max-index}(v) = t\right]dt \end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 it^{i-1} \sum_k k \Pr\left[deg_T(v) = k \mid deg_G(v) = i, \text{max-index}(v) = t\right] dt \\
&= \int_0^1 it^{i-1} E\left[deg_T(v) \mid deg_G(v) = i, \text{max-index}(v) = t\right] dt \\
&\approx \int_0^1 it^{i-1}(i-1)t^3 dt \;=\; i(i-1)\int_0^1 t^{i-1+3} dt \;=\; \frac{i(i-1)}{i+3}. \qquad (11)
\end{aligned}
$$

This completes the Theorem. ■

Note: For high-degree nodes $\frac{i}{i+3} \geq \frac{6}{7}$, so for high-degree nodes we have that

$$E\left[deg_T(v) \mid deg_G(v) = i\right] \geq \frac{6}{7}(i-1) \qquad (12)$$

Theorem 4 *Let $m = \frac{i(i-1)}{2(i+3)}$. Then for every node v it holds that*

$$\Pr\left[deg_T(v) \geq m \mid deg_G(v) = i\right] \geq 1 - \epsilon(i),$$

where $\epsilon(i) = e^{-\frac{i(i-1)}{8(i+3)}}$.

Proof: Recall that the events that each of v's copies give rise to a visible edge are approximately independent. Therefore we can use the Chernoff bound (cf. [MR90]). Let $E = E\left[deg_T(v) \mid deg_G(v) = i\right]$. Then using Theorem 3, we get

$$\Pr\left[deg_T(v) < m \mid deg_G(v) = i\right] < \Pr\left[deg_T(v) < E/2 \mid deg_G(v) = i\right] < e^{-E/8} \approx e^{-\frac{i(i-1)}{8(i+3)}}.$$

■

Note: For high-degree nodes ($i \geq 18$) we have that $\epsilon(i) \geq 0.16$, and for $i \geq 32$ we have that $\epsilon(i) \geq 0.03$.

Our main Theorem is now a corollary of Theorem 4.

Proof of Theorem 2: As a result of Theorem 4 we get that w.h.p for high-degree nodes

$$1 + \frac{i(i-1)}{2(i+3)} \leq deg_T(v) \leq i$$

where $i = deg_G(v)$. Since for high-degree nodes $\frac{i(i-1)}{2(i+3)} \geq \frac{6}{7}(i-1)$, we have that w.h.p for high-degree nodes

$$1 + \frac{3}{7}(i-1) \leq deg_T(v) \leq i$$

Therefore w.h.p for high-degree nodes $deg_T(v) \approx 1 + c(i-1)$, where c is a constant s.t $\frac{3}{7} \leq c \leq 1$. Thus w.h.p for high-degree nodes

$$\Pr\left[deg_T(v) = k\right] \approx \Pr\left[deg_G(v) = \frac{k-1}{c} + 1\right] = \left(\frac{k-1}{c} + 1\right)^{-\gamma} \propto k^{-\gamma}. \qquad (13)$$

■

4 A More Rigorous Analysis

Our analysis of the bias, and especially Lemma 1, used a somewhat cavalier polynomial approximation. In this section we give an alternative derivation of the conservation of the power law tail behavior without relying on the polynomial approximation of the sum. We use a more rigorous approach, but we show a weaker result—that validates the approximations up to multiplicative constants for large k.

Lemma 5 *If* $\gamma > 2$ *then* $Ct \leq C \sum_k k^{1-\gamma} t^k \leq \mu$, *for all* $0 \leq t \leq 1$.

Proof: All summands are positive, so the sum is larger than the first summand. Also the sum is increasing with t and equals μ for $t = 1$. ▮

Lemma 6 $P_{vis}(t) \geq \frac{C^2}{\mu^2}$ *for all* $0 \leq t \leq 1$

Proof: Recall that $a_k = C \cdot k^{-\gamma}$. Using equation (1) and Lemma 5 we further approximate P_{vis} as follows.

$$\begin{aligned} P_{vis}(t) &\approx \frac{1}{\sum_j j a_j t^j} \sum_k k a_k t^k \left(\frac{\sum_j j a_j t^j}{\mu t^2} \right)^k \\ &\geq \frac{1}{\mu} \sum_k C \cdot k^{1-\gamma} t^k \left(\frac{Ct}{\mu t^2} \right)^k \\ &\geq \frac{C^2}{\mu^2}. \end{aligned} \tag{14}$$

Theorem 5 *For large enough* k *there exists* $c_1 > 0$ *such that*

$$c_1 k^{1-\lambda} \leq Pr[deg_T(v) \geq k] \leq C k^{1-\lambda}$$

for a random v.

Proof: The upper value follows immediately from the fact that the visible degree is at most the graph degree. By Lemma 6 we have that $P_{vis}(t) \geq C^2/\mu^2$ for all $0 \leq t \leq 1$. Therefore, for a random v it follows that

$$E[deg_T(v)] \geq \frac{C^2}{\mu^2} deg_G(v)$$

and hence,

$$E[deg_G(v) - deg_T(v)] \leq \left(1 - \frac{C^2}{\mu^2}\right) deg_G(v).$$

By the Markov inequality this means that

$$Pr[deg_G(v) - deg_T(v) > \alpha] < \frac{\left(1 - \frac{C^2}{\mu^2}\right) deg_G(v)}{\alpha}.$$

Take $\alpha = \left(1 - \frac{C^2}{\mu^2} + \epsilon\right) deg_G(v)$ for some constant ϵ. The probability of a node v to have such a difference between its tree-degree and its graph-degree is at most some constant less than 1, and therefore, a constant

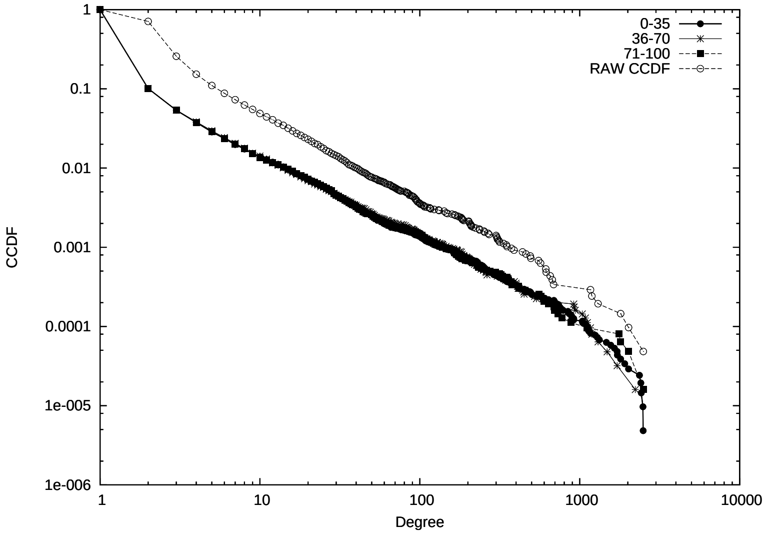

Figure 1: CCDF graphs for BFS trees and raw DIMES data

Data Source	γ
Group 1	2.101
Group 2	2.079
Group 3	2.072
Raw DIMES Data	2.126

Table 1: Sampled power law exponent γ for each group.

fraction of the nodes have a degree proportional to the original degree. Therefore, the tail of the distribution has a power law with exponent at least γ. ■

Notice that for all $\gamma' < \gamma$, there exists some large K_*, such that $C'K_*^{\gamma'} > CK_*^{\gamma}$. Therefore, the exponent of the power law can not decrease throughout the entire degree sequence.

In fact, since the high degree nodes are discovered almost surely at $t \approx 1$, we expect to see $P_{vis} \approx 1$ for these nodes, and therefore, the behavior of the tail is almost unchanged. Giving an exact bound near $t = 1$ is deferred to a future work.

5 Simulation Results

To further validate our analysis, we conducted a simulation study. We used the data collected by Shavitt and Shir [SS05] in the DIMES project as our underlying graph.

To test whether the choice of the BFS tree root has a noticeable effect on the resulting degree distribution, the graph vertices were split into the following 3 groups, based on their graph degree:

1. Low-degree nodes: $1 \leq deg_G(v) < 35$,

2. Medium-degree nodes: $36 \leq deg_G(v) < 70$,

3. High-degree nodes: $deg_G(v) \geq 71$.

From each group we selected 10 nodes at random and constructed a BFS tree for each, where the selected node was a tree root. An average CCDF[2] of degree distribution was then calculated for every group. We compared the resulting curves to the original connectivity data, collected by DIMES.

Figure 1 shows the plotted CCDF curves for the three groups, and the curve for the raw DIMES data. The figure clearly shows the familiar power-law curves in all cases, and we can see that the curves are almost parallel graphs, indicating a similar value of the power-law exponent γ.

Table 1 contains the computed values of γ for each group. We can immediately see that the values of γ on the sampled trees (2.072–2.101) are very close to the true power-law exponent (2.126), thus validating our analysis that the bias is minor. Furthermore, Table 1 shows that the γ values for the three groups are all close to one another, with a minor decrease in value as the degree of the root grows. Thus, it seems that the value of the power-law exponent in the sampled tree is largely invariant to the degree of the tree root.

6 Conclusions and Future Work

We have shown that if the underlying graph degree distribution obeys a power-law with an exponent $\gamma > 2$ (as is the case in the AS-graph) then with w.h.p the degree distribution of the high-degree vertices of the sampled graph also follows a power-law, with the same exponent value. Therefore, the bias observed in tree-sampling regarding the degree distribution of the sampled graph is not significant under these conditions. Furthermore, since according the non-tree-sampled data of [SS05] the AS-graph degree distribution *does* obey a power-law with an exponent γ between 2 and 3, we conclude that the bias observed in the degree distribution of the BGP data is negligible. Thus, the commonly held view of the Internet's topology as having a degree distribution of a power-law form with an exponent $2 < \gamma < 3$ seems to be correct, and unlikely to be a by-product of the BGP data collection process.

Acknowledgment: we thank Sagy Bar for producing the CCDF graphs from the DIMES data.

References

[AB00] Réka Albert and Albert-László Barabási. Topology of evolving networks: Local events and universality. *Physical Review Letters.*, 85(24):5234–5237, December 2000.

[ACKM05] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling. In *Symposium on the Theory of Computing (STOC)'2005*, Baltimore, MD, USA, May 2005.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 15 October 1999.

[BB01] G. Bianconi and A. L. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436–442, 2001.

[BBBC01] P. Barford, A. Bestavros, J. Byers, and M. Crovella. On the marginal utility of network topology measurements. In *Proc. ACM SIGCOMM*, 2001.

[BGW04] S. Bar, M. Gonen, and A. Wool. An incremental super-linear preferential Internet topology model. In *Proc. 5th Annual Passive & Active Measurement Workshop (PAM), LNCS 3015*, pages 53–62, Antibes Juan-les-Pins, France, April 2004. Springer-Verlag.

[2] Cumulative Complementary Distribution Function.

[BGW05] S. Bar, M. Gonen, and A. Wool. A geographic directed preferential Internet topology model. In *Proc. 13th IEEE Symp. Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 325–328, Atlanta, GA, September 2005.

[Bol85] B. Bollobás. *Random Graphs*. Academic Press, 1985.

[BS02] R.X. Brunet and I.M. Sokolov. Evolving networks with disadvantaged long-range connections. *Physical Review E*, 66(026118), 2002.

[BSH03] Tony Bates, Philip Smith, and Geoff Huston. Cidr reports, 2003. `http://www.cidr-report.org/`.

[BT02] T. Bu and D. Towsley. On distinguishing between Internet power-law generators. In *Proc. IEEE INFOCOM'02*, New-York, NY, USA, April 2002.

[CCG$^+$02] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. In *Proc. IEEE INFOCOM'02*, New-York, NY, USA, April 2002.

[CM04a] A. Clauset and C. Moore. Traceroute sampling makes random graphs appear to have power law degree distributions. 2004. Preprint, Submitted to Physical Review Letters.

[CM04b] A. Clauset and C. Moore. Why mapping the Internet is hard, 2004. arXiv:cond-mat/0407339.

[ER60] P. Erdős and A. Renyi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 5:17–61, 1960.

[FFF99] C. Faloutsos, M. Faloutsos, and P. Faloutsos. On power-law relationships of the Internet topology. In *Proc. of ACM SIGCOMM'99*, pages 251–260, August 1999.

[GT00] R. Govindan and H. Tangmunarunki. Heuristics for Internet map discovery. In *Proc. IEEE INFOCOM'00*, pages 1371–1380, Tel-Aviv, Israel, March 2000.

[KRR01] P.L. Krapivsky, G.J. Rodgers, and S. Render. Degree distributions of growing networks. *Physical Review Letters*, 86:5401, 2001.

[LBCX03] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurments. In *Proc. IEEE INFOCOM'03*, 2003.

[LC03] X. Li and G. Chen. A local-world evolving network model. *Physica A*, 328:274–286, 2003.

[MR90] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1990.

[PR04] T. Petermann and P. De Los Rios. Exploration of scale-free networks - do we measure the real exponents? *Europhysics Letters*, 38:201–204, 2004.

[RN04] H. Reittu and I. Norros. On the power law random graph model of the Internet. *Performance Evaluation*, 55, January 2004.

[Rou05] University of Oregon Route Views project, 2005. `http://www.routeviews.org`.

[SS05] Y. Shavitt and E. Shir. DIMES: Let the Internet measure itself, 2005. `http://www.arxiv.org/abs/cs.NI/0506099`.

[TGJ$^+$02] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. Network topology generators: Degree based vs. structural. In *Proc. ACM SIGCOMM*, 2002.

[TPSF01] L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for Internet topology. In *IEEE Global Internet*, San Antonio, TX, November 2001.

[WGJ$^+$02] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. Scaling phenomena in the Internet: Critically examining criticality. *Proceedings of the National Academy of Sciences of the United States of America*, 99:2573–2580, February 2002.

[WJ02] Jared Winick and Sugih Jamin. Inet-3.0: Internet topology generator. Technical Report UM-CSE-TR-456-02, Department of EECS, University of Michigan, 2002.