On the Bias of Traceroute Sampling

or, Power-law Degree Distributions in Regular Graphs

Dimitris Achlioptas
Microsoft Research
Microsoft Corporation
Redmond, WA 98052

optas@microsoft.com

David Kempe
Department of Computer Science
University of Southern California
Los Angeles, CA 90089

dkempe@usc.edu

Aaron Clauset
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131

aaron@cs.unm.edu

Cristopher Moore
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131

moore@cs.unm.edu

ABSTRACT

Understanding the structure of the Internet graph is a crucial step for building accurate network models and designing efficient algorithms for Internet applications. Yet, obtaining its graph structure is a surprisingly difficult task, as edges cannot be explicitly queried. Instead, empirical studies rely on *traceroutes* to build what are essentially single-source, all-destinations, shortest-path trees. These trees only sample a fraction of the network's edges, and a recent paper by Lakhina et al. found empirically that the resuting sample is intrinsically biased. For instance, the observed degree distribution under traceroute sampling exhibits a power law even when the underlying degree distribution is Poisson.

In this paper, we study the bias of traceroute sampling systematically, and, for a very general class of underlying degree distributions, calculate the likely observed distributions explicitly. To do this, we use a continuous-time realization of the process of exposing the BFS tree of a random graph with a given degree distribution, calculate the expected degree distribution of the tree, and show that it is sharply concentrated. As example applications of our machinery, we show how traceroute sampling finds power-law degree distributions in both δ-regular and Poisson-distributed random graphs. Thus, our work puts the observations of Lakhina et al. on a rigorous footing, and extends them to nearly arbitrary degree distributions.

Categories and Subject Descriptors: F.2.2, G.2.1

General Terms: Theory

Keywords: Internet mapping, power laws, random graphs

1. INTRODUCTION

Owing to the great importance of the Internet as a medium for communication, a large body of recent work has focused on its topological properties. Perhaps most famously, Faloutsos et al. [9] exhibited a power-law degree distribution in the Internet graph at the router level (i.e., the level at which the Internet Protocol (IP) operates). Similar results were obtained in [10, 2]. Based on these and other topological studies, it is widely believed that the Internet's degree distribution has a power-law form with exponent $2 < \alpha < 3$, i.e., the fraction a_k of vertices with degree k is proportional to $k^{-\alpha}$. These results have motivated both the search for natural graph growth models that give similar degree distributions (see for instance [8]) and research into the question of how the topology might affect the performance of Internet algorithms and mechanisms (for instance [17]).

However, unlike graphs such as the World Wide Web [13] in which links from each site can be readily observed, the physical connections between routers on the Internet cannot be queried directly. Without explicitly knowing which routers are connected, how can one obtain an accurate map of the Internet? Internet mapping studies typically address this issue by sampling the network's topology using *traceroutes*: packets are sent across the network in such a way that their paths are annotated with the IP addresses of the routers that forward them. The union of many such paths then forms a partial map of the Internet. While actual routing decisions involve multiple protocols and network layers, it is a common assumption that the packets follow shortest paths between their source and destination, and recent studies show that this is not far from the truth [15].

Most studies, including the one on which [9] is based, infer the Internet's topology from the union of traceroutes from a single *root* computer to a larger number of (or all) other computers in the network. If each edge has unit cost plus a small random term, the union of these shortest paths is a BFS tree.[1] This model of the sampling process is admittedly

---

[1]Several studies, including [2, 21], have used traceroutes from multiple sources. However, the number of sources used is quite small (to our knowledge, at most 12).

an idealization for several reasons. First of all, most empirical studies only use a subset of the valid IP addresses as destinations. Secondly, for technical reasons, some routers may not respond to traceroute queries. Thirdly, a single router may annotate different traceroutes with different IP addresses, a problem known as *aliasing*. These issues are known to introduce noise into the measured topology [1, 5].

However, as Lakhina et al. [14] recently pointed out, traceroute sampling has a more fundamental bias, one which is well-captured by the BFS idealization. Specifically, in using such a sample to represent the network, one tacitly assumes that the sampling process is unbiased with respect to the parameters under consideration, such as node degrees. However, an edge is much more likely to be *visible*, i.e., included in the BFS tree, if it is close to the root. Moreover, since in a random graph, high-degree vertices are more likely to be encountered early on in the BFS tree, they are sampled more accurately than low-degree vertices. Indeed, [14] showed empirically that for Erdős-Rényi random graphs $G(n, p)$ [7], which have a Poisson degree distribution, the observed degree distribution under traceroute sampling follows a power law, and this has been verified analytically by Clauset and Moore [6]. In other words, the bias introduced by traceroute sampling can make power laws appear where none existed in the underlying graph! Even when the underlying graph actually does have a power-law degree distribution $k^{-\alpha}$, Petermann and De Los Rios [22] and Clauset and Moore [6] showed numerically that traceroute sampling can significantly underestimate its exponent α.

This inherent bias in traceroute sampling (along with the fact that no alternatives are technologically feasible at this point) raises the following interesting question: Given the true degree distribution $\{a_k\}$ of the underlying graph, can we predict the degree distribution that will be observed after traceroute sampling? Or, in pure graph-theoretic terms: can we characterize the degree distribution of a BFS tree for a random graph with a given degree distribution?

Our answers to these questions quantify precisely the bias introduced by traceroute sampling, while verifying formally the empirical observations of Lakhina et al. [14]. In addition, they can be considered as a significant first step toward a much more ambitious and ultimately more practical goal of inferring the true underlying distribution of the Internet from the biased observation.

Our Results

Our main result in this paper is Theorem 2, which explicitly characterizes the observed degree distribution as a function of the true underlying distribution, to within sharp concentration. When we say that $\{a_k\}$ is a degree distribution, what we mean precisely is that the graph contains $a_k \cdot n$ nodes of degree k. In proving the result, we restrict our attention to underlying distributions which are "not too heavy-tailed," and in which all nodes have degree at least 3:

DEFINITION 1. *A degree distribution $\{a_k\}$ is reasonable if $a_k = 0$ for $k < 3$, and there exist constants $\alpha > 2$ and $C > 0$ such that $a_k < C \cdot k^{-\alpha}$ for all k.*

The requirement that the degree distribution be bounded by a power law $k^{-\alpha}$ with $\alpha > 2$ is made mostly for technical convenience. Among other things, it implies that the mean degree $\delta = \sum_k k a_k$ of the graph is finite (although

the variance is infinite for $\alpha \le 3$). Note that this requirement is consistent with the conjectured range $2 \le \alpha \le 3$ for the Internet [9, 10]. The requirement that the minimum degree be at least 3 implies, through a simple counting argument, that the graph is w.h.p. connected.[2] This is convenient since it ensures that the breadth-first tree reaches the entire graph. However, as we discuss below, this requirement can be relaxed, and in the case of disconnected graphs such as $G(n, p = \delta/n)$, we can indeed analyze the breadth-first tree built on the giant component.

In order to speak precisely about a random (multi)graph with a given degree sequence, we will use the *configuration model* [3]: 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. Our main result can then be stated as follows:

THEOREM 2. *Let $\{a_j\}$ be a reasonable degree sequence. Let G be a random multigraph with degree distribution $\{a_j\}$, and assume that G is connected. Let T be a breadth-first tree on G, and let A_j^{obs} be the number of vertices of degree j in T. Then, there exists a constant $\zeta > 0$ such that with high probability, $\left| A_j^{\mathrm{obs}} - a_j^{\mathrm{obs}} n \right| < n^{1-\zeta}$ for all j, where*

$$a_{m+1}^{\mathrm{obs}} = \sum_i a_i \left[\int_0^1 i t^{i-1} \binom{i-1}{m} \right.$$
$$\left. p_{\mathrm{vis}}(t)^m \left(1 - p_{\mathrm{vis}}(t)\right)^{i-1-m} \, \mathrm{d}t \right] ,$$

$$p_{\mathrm{vis}}(t) = \frac{1}{\sum_j j a_j t^j} \sum_k k a_k t^k \left(\frac{\sum_j j a_j t^j}{\delta t^2} \right)^k .$$

We can use the notion of generating functions [25] to obtain a more concise expression of all a_{m+1}^{obs} as follows: if $g(z) = \sum_{j=0}^\infty a_j z^j$, then a_j^{obs} is the coefficient of z^j in

$$g^{\mathrm{obs}}(z) = z \int_0^1 g' \left[t - \frac{(1-z)}{g'(1)} g' \left(\frac{g'(t)}{g'(1)} \right) \right] \, \mathrm{d}t . \qquad (1)$$

The bulk of this paper, namely Sections 2–5, is devoted to the proof of Theorem 2. In Section 6, we apply our general result to δ-regular graphs and graphs with Poisson degree distributions. In both cases, we find that the observed degree distribution follows a power law $k^{-\alpha}$ with exponent $\alpha = 1$. In the case of Poisson degree distributions, our work thus subsumes the work of Clauset and Moore [6].

The proof of this result is based on a process which gradually discovers the BFS tree (see Section 2). By mapping it to a continuous-time process analogous to Kim's Poisson cloning model [12], we can avoid explicitly tracking the (rather complicated) state of the FIFO queue that arises in the process, and in particular the complex relationship between a vertex's degree and its position in the queue. This allows us to calculate the expected degree distribution to within $o(1)$ in Section 3. In Section 5, we see how these calculations can be rephrased in terms of generating functions, to yield the alternate formulation of Theorem 2. The

---

[2]We say that a sequence of events $\mathcal{E}_n$ occurs *with high probability (w.h.p.)* if $\mathrm{Prob}[\mathcal{E}_n] = 1 - o(1)$ as $n \to \infty$, and *with overwhelmingly high probability (w.o.h.p.)* if $\mathrm{Prob}[\mathcal{E}_n] = 1 - o(n^{-c})$ for all c. Note that by the union bound, the conjunction of a polynomial number of events, each of which occurs w.o.h.p., occurs w.o.h.p.

concentration part of the result, in Section 4, analyzes a different, and much more coarse-grained, view of the process. By carefully conditioning on the history of the process, we can apply a small number of Martingale-style bounds to obtain overall concentration.

2. A CONTINUOUS-TIME PROCESS

2.1 Breadth-First Search

We can think of the breadth-first tree as being built one vertex at a time by an algorithm that explores the graph. At each step, every vertex in the graph is labeled *explored*, *untouched*, or *pending*. A vertex is explored if both it and its neighbors are in the tree; untouched if it is still outside the tree; and pending if it is on the boundary of the tree, i.e., it may still have untouched neighbors. Pending vertices are kept in a queue $\mathcal{Q}$, so that they are explored in first-in, first-out order. The process is initialized by labeling the root vertex pending, and all other vertices untouched. Whenever a pending vertex is popped from $\mathcal{Q}$ and explored, all of its currently untouched neighbors are appended to $\mathcal{Q}$, and the connecting edges are *visible*. On the other hand, edges to neighbors that are already in the queue are not visible.

For the analysis in this paper, it is convenient to think of the algorithm as exploring the graph one copy at a time, instead of one node at a time. The queue will then contain copies instead of vertices. At each step, the partner v of the copy u at the head of the queue is *exposed*, and both of them are removed from the matching. Also, all of v's siblings are added to the queue, unless they were in the queue already. (We refer to two copies of the same vertex as *siblings*.) We will say that an unexposed copy is *enqueued* if it is in $\mathcal{Q}$, and *untouched* if it is not. Thus, a copy is untouched if its vertex is, and enqueued if its vertex is pending and the edge incident to it has yet to be explored. Formally, the breadth-first search then looks as follows:

Algorithm 1 Breadth-First Search at the Copy Level

1: **while** $\mathcal{Q}$ is nonempty **do**
2:    Pop a copy u from the head of $\mathcal{Q}$
3:    Expose u's partner v
4:    **if** v is untouched **then**
5:      Add the edge (u, v) to T
6:      Append v's siblings to $\mathcal{Q}$
7:    **else**
8:      Remove v from $\mathcal{Q}$
9:    **end if**
10: **end while**

An edge will be visible and included in T if, at the time one of its endpoints reaches the head of the queue, the other endpoint is still untouched.

2.2 Exposure on the fly

Because G is a uniformly random multigraph conditioned on its degree sequence, the matching on the copies is *uniformly random*. By the principle of deferred decisions [20], we can define this matching "on the fly," choosing u's partner v uniformly at random from among all the unexposed copies at the time.

One way to make this random choice is as follows. At the outset, each copy is given a real-valued index x chosen uniformly at random from the unit interval $[0, 1]$. Then, at each step, u's partner v is chosen as the unexposed copy with the the largest index. Thus, it is convenient to think of the algorithm as taking place in continuous time, where t decreases from 1 to 0: at time t, the copy at the head of the queue is matched with the unexposed copy of index t. Since the indices of v's siblings are uniformly random, while conditioned on being less than t, this approach maintains the following powerful kind of uniform randomness: at time t, the indices of the unexposed copies, both inside and outside the queue, are uniformly random in $[0, t)$.

We define the *maximum index* of a vertex to be the maximum of all its copies' indices. At any time t, the untouched vertices are precisely those whose maximum index is less than t, and the explored or pending vertices (whose copies are explored or enqueued) are those whose maximum index is greater than t. This observation allows us to carry out an explicit analysis without having to track the (rather complicated) state of the system as a function of time.

At a given time t, let $C_{\mathrm{unex}}(t)$ and $C_{\mathrm{unto}}(t)$ denote the number of unexposed and untouched copies, and let $V_{\mathrm{unto},j}(t)$ denote the number of untouched vertices of degree j; note that $C_{\mathrm{unto}}(t) = \sum_j j V_{\mathrm{unto},j}(t)$. We start by calculating the expectation of these quantities. The probability that a vertex of degree j has maximum index less than t is exactly t^j; therefore, $\mathrm{E}[V_{\mathrm{unto},j}(t)] = a_j t^j n$, and

$$\mathrm{E}[C_{\mathrm{unto}}(t)] \quad = \quad \sum_j j a_j t^j n \quad =: \quad c_{\mathrm{unto}}(t) \cdot n \ . \qquad (2)$$

To calculate $\mathrm{E}[C_{\mathrm{unex}}(t)]$, recall that the copy at the head of the queue has a uniformly random index conditioned on being less than t. Therefore, the process forms a matching on the list of indices as follows: take the indices in decreasing order from 1 to 0, and at time t match the index t with a randomly chosen index less than t. This creates a uniformly random matching on the δn indices. Now, note that a given index is still remaining at time t if both it and its partner are less than t, and since the indices are uniformly random in $[0, 1]$ the probability of this is t^2. Thus, the expected number of indices remaining at time t is

$$\mathrm{E}[C_{\mathrm{unex}}(t)] \quad = \quad \delta t^2 n \quad =: \quad c_{\mathrm{unex}}(t) \cdot n \ . \qquad (3)$$

The following lemma shows that $V_{\mathrm{unto},j}(t)$, $C_{\mathrm{unto}}(t)$ and $C_{\mathrm{unex}}(t)$ are concentrated within $o(n)$ of their expectations throughout the process. Note that we assume here that the graph G is connected, since otherwise, the process is not well-defined for all $t \in [0, 1]$.

LEMMA 3. *Let $\{a_j\}$ be a reasonable degree distribution, and assume that G is connected. Then, for any constants $\beta < \min(\frac{1}{2}, \frac{\alpha-2}{2})$ and $\epsilon > 0$, the following hold simultaneously for all $t \in [0, 1]$ and for all $j < n$, w.o.h.p.:*

$$\left| V_{\mathrm{unto},j}(t) - a_j t^j \cdot n \right| < n^{1/2+\epsilon}$$

$$|C_{\mathrm{unex}}(t) - c_{\mathrm{unex}}(t) \cdot n| < n^{1/2+\epsilon}$$

$$|C_{\mathrm{unto}}(t) - c_{\mathrm{unto}}(t) \cdot n| < n^{1-\beta},$$

where $c_{\mathrm{unto}}(t)$ and $c_{\mathrm{unex}}(t)$ are given by (2), (3).

Note that this concentration becomes weaker as $\alpha \to 2$, since then $\beta \to 0$.

Proof. Our proof is based on the following form of the Hoeffding Bound [11, 16]:

THEOREM 4 (THEOREM 3 FROM [16]). *If $X_1, \ldots, X_k$ are independent, non-negative random variables with $X_i \leq b_i$ for all i, and $X = \sum_i X_i$, then for any $\Delta \geq 0$:*

$$\text{Prob}[|X - \text{E}[X]| \geq \Delta] \leq 2\text{e}^{-2\Delta^2/\sum_i b_i^2} \ .$$

First, $V_{\text{unto},j}(t)$ is a binomial random variable distributed as $\text{Bin}(a_j n, t^j)$. By applying Theorem 4 to $a_j n$ variables bounded by 1, the probability that $V_{\text{unto},j}(t)$ differs by $\Delta = n^{1/2+\epsilon}$ from its expectation is at most $2\text{e}^{-2n^{2\epsilon}/a_j} \leq \text{e}^{-n^{2\epsilon}}$. Thus, at each individual time t and for each j, the stated bound on $V_{\text{unto},j}(t)$ holds w.o.h.p.

We wish to show that this bound holds w.o.h.p. for all j and all t, i.e., that the probability that it is violated for any t and any j is $o(n^{-c})$ for all c. Notice that the space of all times t is infinite, so we cannot take a simple union bound. Instead, we divide the interval $[0, 1]$ into sufficiently small discrete subintervals, and take a union bound of those. Let $m = \sum_j j a_j n = \delta n$ be the total number of copies, where δ is the mean degree (recall that δ is finite, because $\{a_j\}$ is reasonable). We divide the unit interval $[0, 1]$ into m^b intervals of size m^{-b}, where b will be set below. By a union bound over the $\binom{m}{2}$ pairs of copies, with probability at least $1 - m^{2-b}$, each interval contains the index of at most one copy, and therefore at most one event of the queue process. Conditioning on this event, $V_{\text{unto},j}(t)$ changes by at most 1 during each interval, so if $V_{\text{unto},j}(t)$ is close to its expectation at the boundaries of each interval, it is close to its expectation for all $t \in [0, 1]$. In addition, we take a union bound over all j. The probability that the stated bound is violated for any j in any interval is then at most

$$n\left(m^b \, \text{e}^{-n^{2\epsilon}} + m^{2-b}\right) = O(n^{3-b}) \ ,$$

which is $o(n^{-c})$ if $b > c + 3$.

For the concentration of $C_{\text{unex}}(t)$, we notice that unexposed copies come in matched pairs, both of which have index less than t. Therefore, $C_{\text{unex}}(t)$ is twice a binomial random variable distributed as $\text{Bin}(\sum_j j a_j n/2, t^2)$. Applying Theorem 4 with $\Delta = n^{1/2+\epsilon}$ gives the result for fixed t, and taking a union bound over t as in the previous paragraph shows the concentration of $C_{\text{unex}}(t)$.

To prove concentration of $C_{\text{unto}}(t)$ for fixed t, we let X_i be the number of copies of node i that are untouched at time t. Then, $C_{\text{unto}}(t) = \sum_i X_i$, and the denominator in the exponent for the bound of Theorem 4 is

$$\sum_i b_i^2 = \sum_j j^2 a_j n < Cn \sum_j j^{2-\alpha} < \begin{cases} O(n^{4-\alpha}) & \alpha < 3 \\ O(n \log n) & \alpha = 3 \\ O(n) & \alpha > 3 \end{cases}$$

Hence, by Theorem 4, whenever $\beta < \min(\frac{1}{2}, \frac{\alpha-2}{2})$, we obtain that $|C_{\text{unto}}(t) - \text{E}[C_{\text{unto}}(t)]| \leq n^{1-\beta}$ w.o.h.p. A union bound over t as before completes the proof. ∎

3. EXPECTED DEGREE DISTRIBUTION

In this section, we begin the proof of Theorem 2 by analyzing the continuous-time process defined in Section 2, and calculating the expected degree distribution of the tree T.

By linearity of expectation, the expected number of vertices of degree j in T is the sum, over all vertices v, of the probability that j of v's edges are visible. Consider a given vertex v of degree i. It is touched when its copy with maximum index is matched to the head of the queue, at which

time its $i - 1$ other copies join the tail of the queue. If m of these give rise to visible edges, then v's degree in T will be $m + 1$, namely these m outgoing edges plus the edge connecting v back toward the root of the tree.

Let $\rho_{i,m}$ denote the probability of this event, i.e., that a vertex of degree i has m copies that give rise to visible edges. Then the expected degree distribution is given by

$$\text{E}\left[A_{m+1}^{\text{obs}}\right] = n \sum_i a_i \rho_{i,m} \ . \tag{4}$$

Moreover, let $\rho_{i,m}(t)$ denote the probability of this event given that v has maximum index t. Then, since t is the maximum of i independent uniform variables in $[0, 1]$, its probability distribution is $\text{d}t^i/\text{d}t = it^{i-1}$, and we have

$$\rho_{i,m} = \int_0^1 it^{i-1} \rho_{i,m}(t) \, \text{d}t \ . \tag{5}$$

Our goal is then to calculate $\rho_{i,m}(t)$.

Let us start by calculating the probability $P_{\text{vis}}(t)$ that, if v has index t, a given copy of v other than the copy with index t—that is, a given copy which is added to the queue at time t—gives rise to a visible edge. Call this copy u, and call its partner w. According to Algorithm 1, the edge (u, w) is visible if and only if (1) u makes it to the head of the queue without being matched first, and (2) when it does, w is still untouched. But (1) is equivalent to saying that w is untouched at time t, since if w is already in the queue at time t, it is ahead of u, and u will be matched before it reaches the head of the queue. Similarly, (2) is equivalent to saying that *all of w's siblings' partners* are untouched at time t, since if any of these are already in the queue at time t, and thus ahead of u, then w's vertex will be touched, and w enqueued, by the time u reaches the head of the queue.

Given the number of untouched and unexposed copies $C_{\text{unto}}(t)$ and $C_{\text{unex}}(t)$ at the time t when u joins the queue, the probability that its uniformly random partner w is untouched is $P_{\text{unto}}(t) = C_{\text{unto}}(t)/C_{\text{unex}}(t)$. Conditioning on this event, the probability that w belongs to a vertex with degree k is $P_{\text{unto},k}(t) = kV_{\text{unto},k}(t)/C_{\text{unto}}(t)$. We require that the partners of w's $k - 1$ siblings are also untouched. If we ignore the fact that we are choosing untouched copies without replacement (and that one untouched copy has already been taken for w), and if we assume that v, its neighbors, and its neighbors' neighbors form a tree (i.e., that v does not occur in a triangle or 4-cycle, and that neither it nor its neighbors have any multiple edges), then the probability that these $k - 1$ copies are all untouched is $P_{\text{unto}}(t)^{k-1}$. This gives

$$P_{\text{vis}}(t) = P_{\text{unto}}(t) \sum_k P_{\text{unto},k}(t) \, P_{\text{unto}}(t)^{k-1}$$
$$= \sum_k P_{\text{unto},k}(t) \, P_{\text{unto}}(t)^k. \tag{6}$$

Since $V_{\text{unto},k}(t)$, $C_{\text{unto}}(t)$ and $C_{\text{unex}}(t)$ are concentrated according to Lemma 3, substituting their expectations then gives a good approximation for $P_{\text{vis}}(t)$, namely

$$p_{\text{vis}}(t) = \sum_k \frac{ka_k t^k}{c_{\text{unto}}(t)} \left(\frac{c_{\text{unto}}(t)}{c_{\text{unex}}(t)}\right)^k \ . \tag{7}$$

Then, if we neglect the possibility of self-loops and parallel edges involving u and its siblings, and again ignore the fact

that we are choosing without replacement (i.e., that processing each sibling changes C_{unto}, C_{unex}, and P_{vis} slightly) the events that each of u's siblings give rise to a visible edge are independent, and the number m of visible edges is approximately binomially distributed as $\text{Bin}(i-1, p_{\text{vis}}(t))$.

We wish to confirm this analysis by showing that w.h.p. v, its neighbors, and its neighbors' neighbors form a tree. It is easy to show this for graphs with bounded degree; however, for power-law degree distributions $a_k \sim k^{-\alpha}$, it is somewhat delicate, especially for α close to 2. The following lemmas show that there are very few vertices of very high degree, and then show that the above is w.h.p. true of v if v has sufficiently low degree. We then show that we can think of all the copies involved as chosen with replacement. Recall that the mean degree $\delta = \sum_j j a_j$ is finite, and let $\beta < \min(\frac{1}{2}, \frac{(\alpha-2)}{2})$ as in Lemma 3.

LEMMA 5. *The probability that a random copy belongs to a vertex of degree greater than k is $o(k^{-2\beta})$.*

Proof.   This probability is

$$\frac{\sum_{j>k} j a_j}{\sum_j j a_j} < \frac{C}{\delta} \sum_{j>k} j^{1-\alpha} < \frac{C}{\delta(2-\alpha)} k^{-(\alpha-2)} = o(k^{-2\beta}) \ .$$

∎

LEMMA 6. *There are constants $\gamma > \eta > 0$ such that if v is a vertex of degree $i < n^\eta$, then the probability that v or its neighbors have a self-loop or multiple edge, or that v is part of a triangle or a cycle of length 4, is $o(n^{-\gamma})$. Thus, v, its neighbors, and its neighbors' neighbors form a tree with probability $1 - o(n^{-\gamma})$.*

Proof.   First, we employ Lemma 5 to condition on the event that none of v's neighbors have degree greater than n^λ, where λ (and η) will be determined below. By a union bound over these $i < n^\eta$ neighbors, this holds with probability $1 - o(n^{\eta-2\lambda\beta})$. (Unfortunately, we cannot also condition on v's neighbors' neighbors having degree at most n^λ without breaking this union bound.)

Now, if we choose two copies independently and uniformly at random, the probability that they are both copies of a given vertex of degree $j < n^\lambda$ is $j(j-1)/(\delta n)^2 < n^{2\lambda-2}$, and the probability that they are both copies of *any* such vertex is at most $n^{2\lambda-1}$. Moreover, the probability that two random copies are siblings, regardless of the degree of their vertex, is

$$P_{\text{sib}} = \frac{\sum_j j(j-1)a_j n}{\left(\sum_j j a_j n\right)^2} < \frac{1}{\delta^2 n} \sum_j j^2 a_j = o(n^{-2\beta}) \ .$$

Taking a union bound over all pairs of copies of v, the probability that v has a multiple edge, i.e., that two of its copies are matched to copies of the same neighboring vertex, is at most $i^2 n^{2\lambda-1} = O(n^{2\eta+2\lambda-1})$, and the probability that v contains a self-loop, i.e., that two of its copies are matched, is $O(i^2/(\delta n)) = O(n^{2\eta-1})$. For each of v's neighbors, the probability of parallel edges involving it is at most $n^{2\lambda} P_{\text{sib}} = o(n^{2\lambda-2\beta})$, and the probability of a self-loop is $O(n^{2\lambda}/(\delta n)) = O(n^{2\lambda-1})$. Taking a union bound over all of v's neighbors, the probability that any of them have a self-loop or multiple edge is $o(n^{\eta+2\lambda-2\beta})$.

To determine the expected number of triangles containing v, we notice that any such triangle contains two copies each

from v and two of its neighbors, and edges between the appropriate pairs. A given pair of copies is connected with probability $O(1/(\delta n))$, so the expected number is

$$O\left(n^2 \, n^{2\eta} (n^{2\lambda})^2 / (\delta n)^3\right) = O(n^{2\eta+4\lambda-1}) \ .$$

Similarly, each 4-cycle involves two copies each of v and two of its neighbors, such that one copy from each of the neighbors is matched with one copy of v, and the other two copies are matched with copies of the same node. Thus, the expected number of 4-cycles involving v is

$$O\left(n^2 \, n^{2\eta} (n^{2\lambda})^2 P_{\text{sib}} / (\delta n)^2\right) = o(n^{2\eta+4\lambda-2\beta}) \ .$$

Collecting all these events, the probability that the statement of the lemma is violated is

$$o(n^{-\gamma}) \text{ where } \gamma = -\max(\eta - 2\lambda\beta, 2\eta + 4\lambda - 2\beta) \ .$$

If we set $\eta = \beta^2/6$ and $\lambda = \beta/4$, then $\gamma = \beta^2/3$.   ∎

The next lemma shows that, conditioning on the event of Lemma 6, the copies discussed in our analysis above can be thought of as chosen with replacement, as long as we are not too close to the end of the process where untouched copies become rare. Therefore, the number of visible edges is binomially distributed.

LEMMA 7. *Let η, γ be defined as in Lemma 6. There exists a constant $\theta > 0$ such that for $t \in [n^{-\theta}, 1]$ and $i < n^{-\eta}$,*

$$|\rho_{i,m}(t) - \text{Prob}[\text{Bin}(i-1, P_{\text{vis}}(t)) = m]| < n^{-\gamma} \ .$$

Proof.   Let $d_{\min}$ be the minimum degree of the graph, i.e., the smallest j such that $a_j > 0$. Note that $d_{\min} \geq 3$, and set $\theta = \beta/(2d_{\min}) < 1/12$. For $t \geq n^{-\theta}$, we have that $\text{E}[C_{\text{unex}}(t)] = \delta t^2 n = \Omega(n^{1-\beta/d_{\min}})$, and this bound holds w.o.h.p. by Lemma 3.

Conditioning on v's neighbors having degree at most n^λ as in Lemma 6, the number of visible edges of v is determined by a total of at most $n^{\eta+\lambda}$ copies. These are chosen without replacement from the unexposed copies. If we instead choose them with replacement, the probability of a collision in which some copy is chosen twice is at most $(n^{\eta+\lambda})^2/C_{\text{unex}}(t) = O(n^{2\eta+2\lambda+\beta/d_{\min}-1}) = o(n^{-1/2})$. This can be absorbed into the probability $o(n^{-\gamma})$ that the statement of Lemma 6 does not hold. If there are no collisions, then we can assume the copies are chosen with replacement, and each of v's $i-1$ outgoing edges is independently visible with probability $P_{\text{vis}}(t)$, as defined in (6).   ∎

It is a simple technical matter, which we relegate to the full version of this paper, to show that $\text{Bin}(i-1, P_{\text{vis}}(t))$ is very close to $\text{Bin}(i-1, p_{\text{vis}}(t))$, establishing the following:

LEMMA 8. *Let $\{a_j\}$ be a reasonable degree distribution and assume that G is connected. There are constants $\theta, \kappa, \eta, \mu > 0$, such that for all $t \in [n^{-\theta}, 1-n^{-\kappa}]$ and all $i < n^\eta$, for sufficiently large n,*

$$|\rho_{i,m}(t) - \text{Prob}[\text{Bin}(i-1, p_{\text{vis}}(t)) = m]| < n^{-\mu} \ ,$$

where $p_{\text{vis}}(t)$ is defined in (7).

Finally, combining Lemma 8 with (4), (5), and (7), if

$$a_{m+1}^{\mathrm{obs}} = \sum_i a_i \left[\int_0^1 it^{i-1} \binom{i-1}{m} \right.$$
$$\left. p_{\mathrm{vis}}(t)^m \left(1 - p_{\mathrm{vis}}(t)\right)^{i-1-m} \, dt \right] , \quad (8)$$

where, combining (7) with (2) and (3),

$$p_{\mathrm{vis}}(t) = \frac{1}{\sum_j ja_jt^j} \sum_k ka_kt^k \left(\frac{\sum_j ja_jt^j}{\delta t^2} \right)^k ,$$

then we have the following lemma.

LEMMA 9. *Let $\{a_i\}$ be a reasonable degree sequence and assume that G is connected. There is a constant $\zeta > 0$ such that for sufficiently large n, for all $j < n$*

$$\left| \mathrm{E}\left[A_j^{\mathrm{obs}} \right] - a_j^{\mathrm{obs}}n \right| < n^{1-\zeta} .$$

Proof. There are three sources of error in our estimate of $\mathrm{E}\left[A_j^{\mathrm{obs}} \right]$ for each j. These are the error $n^{-\mu}$ in $\rho_{i,m}(t)$ given by Lemma 8, and the fact that two types of vertices are not covered by that lemma: those with degree greater than n^η, and those which join the queue at some time $t \notin [n^{-\theta}, 1 - n^{-\kappa}]$. The total error is then at most $n^{1-\mu}$ plus the number of vertices of either of these types. The number of vertices of degree greater than n^η is at most

$$n \sum_{j>n^\eta} a_j < Cn \sum_{j>n^\eta} j^{-\alpha} = O(n^{1-(\alpha-1)\eta}) .$$

The number of vertices that join the queue at a time $t \notin [n^{-\theta}, 1 - n^{-\kappa}]$ is at most the number of copies whose index is outside this interval. This is binomially distributed with mean $n^{1-\theta} + n^{1-\kappa}$, and by the Chernoff bound, this is w.o.h.p. less than $n^{1-\zeta}$ for sufficiently large n for any $\zeta < \min(\theta, \kappa)$. The (exponentially small) probability that this bound is violated can be absorbed into $n^{1-\zeta}$ as well. Setting $\zeta < \min(\mu, (\alpha-1)\eta, \theta, \kappa)$ completes the proof. ∎

4. CONCENTRATION

In this section, we prove that the number A_j^{obs} of nodes of observed degree j is tightly concentrated around its expectation $\mathrm{E}\left[A_j^{\mathrm{obs}} \right]$. Specifically, we prove

THEOREM 10. *There is a constant $\rho > 0$ such that, with overwhelmingly high probability, the following holds simultaneously for all j:*

$$\left| A_j^{\mathrm{obs}} - \mathrm{E}\left[A_j^{\mathrm{obs}} \right] \right| \leq O(n^{1-\rho}) .$$

Proof. In order to prove concentration, the style of analysis in the previous section will not be sufficient. Intuitively, the reason is that changing a single edge in the graph can have a dramatic impact on the resulting BFS tree, and thus on the observed degree of a large number of vertices. As a result, it seems unlikely that A_j^{obs} can be decomposed into a large number of small contributions such that their sum can easily be shown to be concentrated. In particular, this rules out the direct application both of Chernoff-style bounds and of martingale-based inequalities.

There is, however, a sense in which martingale bounds will prove helpful. The key is to decompose the evolution of A_j^{obs} into a small number of "bulk moves," and prove concentration for each one of them. Concretely, assume that the BFS tree has already been exposed up to a certain distance r from the root, and that we know the number of copies in the queue, as well as the number of untouched copies at that point. Since all these copies will be matched uniformly at random, one can use an edge-switching martingale bound to prove that the degree distributions of nodes at distance $r+1$ from the root will be sharply concentrated. In fact, this concentration argument applies to the observed degrees of the neighbors of any "batch" of copies that comprise the queue $\mathcal{Q}(t)$ at some time t.

We will implicitly divide the copies in the graph into such "batches" by specifying a set of *a priori fixed* points in time at which we examine the system. That is, we will approximate A_j^{obs} by the sum of the observed degrees of the neighbors of $\mathcal{Q}(t)$ over these time steps. We will show that each of the terms in the sum is sharply concentrated around its expectation, and then prove that the true expectation of A_j^{obs} is not very far from the expectation of the sum that we consider. For the latter part, it is crucial that most vertices be counted exactly once in the sum; this will follow readily from the concentration given by Lemma 3.

To make the above outline precise, we let $Q(t) := |\mathcal{Q}(t)|$ be the number of copies in the queue at time t, and let

$$q(t) := \mathrm{E}[Q(t)] = (c_{\mathrm{unex}}(t) - c_{\mathrm{unto}}(t)) \cdot n$$

be its expected size. We define a sequence of $r \leq \log^2 n$ times at which we observe the queue and its neighbors. We start with $t_1 = 1$. For each i, we let $t_{i+1} \geq 0$ be maximal such that

$$c_{\mathrm{unex}}(t_i) - c_{\mathrm{unex}}(t_{i+1}) \geq q(t_i)/n + 2n^{-\beta} ,$$

where β is defined as in Lemma 3. Depending (deterministically) on the properties of the real-valued functions c_{unex} and c_{unto}, there may be an $i < \log^2 n$ such that t_{i+1} does not exist, namely when $c_{\mathrm{unex}}(t_i) < 2n^{-\beta}$. If so, we let r be that i; otherwise, we let $r = \log^2 n$.

For each degree j, let $B_j(i)$ denote the number of vertices adjacent to $\mathcal{Q}(t_i)$ whose observed degree is j. Lemma 11 below shows that each $B_j(i)$ is sharply concentrated. However, we want to prove concentration for the overall quantity A_j^{obs}. Using a union bound over all $i = 1, \ldots, r$ and summing up the corresponding $B_j(i)$ will give us concentration for A_j^{obs}, assuming that (1) not too many times t_i are considered, (2) nodes are not double-counted for multiple i, and (3) almost all nodes are considered in some batch i.

For the first point, recall that we explicitly chose $r = O(\log^2 n)$. For the second, observe that whenever

$$C_{\mathrm{unex}}(t_i) - C_{\mathrm{unex}}(t_{i+1}) \geq q(t_i) + 2n^{1-\beta} \geq Q(t_i)$$

for all times i, then all of the $\mathcal{Q}(t_i)$ are disjoint. Each of these bounds holds w.o.h.p. by Lemma 3, and by the union bound, w.o.h.p. they hold simultaneously.

This leaves the third point. Here, we first bound the number of nodes that remain unexposed after time t_r. If the construction terminated prematurely (i.e., $r < \log^2 n$), then the fact that $c_{\mathrm{unex}}(0) = 0$ implies that $c_{\mathrm{unto}}(t_r) < 2n^{-\beta}$, so by Lemma 3, at most $O(n^{1-\beta})$ copies remain unexposed w.o.h.p. On the other hand, when $r = \log^2 n$, we can use the fact that the diameter of a random graph is bounded by $\log^2 n$ with probability at least $1 - n^{-1/2}$, which we prove in the full paper using techniques of Bollobás and Chung

[4]. Even if $C_{\text{unto}}(t_r)$ were $\Omega(n)$ in the remaining case, since this occurs with probability at most $n^{-1/2}$, we have $c_{\text{unto}}(t_r) \cdot n = O(n^{1/2}) = O(n^{1-\beta})$ since $\beta < 1/2$.

Let $\mathcal{E}$ denote the event that $\left|V_{\text{unto},j}(t_i) - a_j t_i^j n\right| \leq n^{1/2+\epsilon}$ and $|Q(t_i) - q(t_i)| \leq 2n^{1-\beta}$ hold simultaneously for all i. By Lemma 3, $\mathcal{E}$ occurs w.o.h.p. In that case, we know that (1) all of the sets $\mathcal{Q}(t_i)$ are disjoint, and (2) the union of all the $\mathcal{Q}(t_i)$ excludes at most $2r \cdot n^{1-\beta} + O(n^{1-\beta}) = \tilde{O}(n^{1-\beta})$ copies total (where $\tilde{O}$ includes polylog(n) factors). Thus, $\left|A_j^{\text{obs}} - \sum_{i=1}^r B_j(i)\right| = \tilde{O}(n^{1-\beta})$ w.o.h.p., which implies that $\left|\mathrm{E}\left[A_j^{\text{obs}}\right] - \sum_{i=1}^r \mathrm{E}[B_j(i)]\right| = \tilde{O}(n^{1-\beta})$, since this difference is deterministically bounded above by n.

By Lemma 11 below and a union bound over all i, there is a $\tau > 0$ such that w.o.h.p. $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$ holds simultaneously for all $i = 1, \ldots, r$ and all j. Hence, by a union bound with the event $\mathcal{E}$, and the triangle inequality, the following holds w.o.h.p.:

$$
\begin{aligned}
\left|A_j^{\text{obs}} - \mathrm{E}\left[A_j^{\text{obs}}\right]\right| &\leq \left|A_j^{\text{obs}} - \sum_{i=1}^r B_j(i)\right| \\
&\quad + \sum_{i=1}^r |B_j(i) - \mathrm{E}[B_j(i)]| \\
&\quad + \left|\mathrm{E}\left[A_j^{\text{obs}}\right] - \sum_{i=1}^r \mathrm{E}[B_j(i)]\right| \\
&= \tilde{O}(n^{1-\beta}) + \tilde{O}(n^{1-\tau}) \\
&= O(n^{1-\rho}) \ .
\end{aligned}
$$

for any $\rho < \min(\beta, \tau)$, completing the proof of Theorem 10. ∎

The concentration for one "batch" of nodes at time t_i is captured by the following lemma.

LEMMA 11. *There is a constant $\tau > 0$ such that, for any fixed i, w.o.h.p., $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$ holds simultaneously for all j.*

Proof. As explained above, the idea for the proof is to apply an edge-exposure Martingale-style argument to the nodes that are adjacent to $\mathcal{Q}(t_i)$. We use the following concentration inequality for random variables on matchings due to Wormald [26, Theorem 2.19]. A *switching* consists of replacing two edges $\{p_1, p_2\}$, $\{p_3, p_4\}$ by $\{p_1, p_3\}$, $\{p_2, p_4\}$.

THEOREM 12. *[26] Let X_k be a random variable defined on uniformly random configurations M, M' of k copies, such that, whenever M and M' differ by only one switching,*

$$\left|X_k(M) - X_k(M')\right| \leq c$$

for some constant c. Then, for any $r > 0$,

$$\mathrm{Prob}[|X_k - \mathrm{E}[X_k]| \geq \Delta] < 2\mathrm{e}^{-\Delta^2/(kc^2)} \ .$$

For fixed values q and $\mathbf{b} = b_1, \ldots, b_n$, let $\mathcal{E}_{q,\mathbf{b}}$ denote the event that $Q(t_i) = q$ and $V_{\text{unto},j}(t_i) = b_j$ for all j. Conditioned on $\mathcal{E}_{q,\mathbf{b}}$, the matching on the $q + \sum_j jb_j$ copies is uniformly random. Since any switching changes the value of $B_j(i)$ by at most 2, Theorem 12 implies that

$$|B_j(i) - \mathrm{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}]| \leq n^{1/2+\epsilon} \qquad (9)$$

holds w.o.h.p. for any $\epsilon > 0$. If we knew the queue size q and the number b_j of untouched nodes of degree j exactly, then we could apply Theorem 12 directly.

In reality, we will certainly not know the precise values of q and $\mathbf{b}$. Therefore, we need to analyze the effect that deviations of these quantities will have on our tail bounds. We do this by showing in Lemma 13 below that the conditional expectations $\mathrm{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}]$ are close to the actual expectations $\mathrm{E}[B_j(i)]$. It follows that concentration around the conditional expectation implies concentration around the actual expectation. Specifically, write

$$I^q := \left[q(t_i) - 2n^{1-\beta}, q(t_i) + 2n^{1-\beta}\right]$$

for the interval of possible queue lengths under consideration, and, for some $0 < \epsilon < 1/2$, write

$$I_j^b := \left[a_j t_i^j n - n^{1/2+\epsilon}, a_j t_i^j n + n^{1/2+\epsilon}\right]$$

for the interval of possible numbers of untouched vertices of degree j, as well as $I^b := I_1^b \times \cdots \times I_n^b$ for the range of all possible combinations of numbers of untouched vertices. Now, let $\mathcal{E}_{\leq}$ be the event that $Q(t_i) \in I^q$ and $V_{\text{unto},j}(t_i) \in I_j^b$ for all j. Notice that $\mathcal{E}_{\leq}$ occurs w.o.h.p. by Lemma 3.

Lemma 13 ensures that whenever $q \in I^q$ and $\mathbf{b} \in I^b$, then the conditional expectation is close to the true expectation, i.e., for some $\tau > 0$,

$$|\mathrm{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathrm{E}[B_j(i)]| = O(n^{1-\tau}) \ .$$

Thus, for all such q and $\mathbf{b}$, combining this with (9) and the triangle inequality gives $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$, so the latter occurs w.o.h.p. Finally, a union bound with the event $\mathcal{E}_{\leq}$ and over all j completes the proof. ∎

The final missing step is a bound relating the conditional expectation of $B_j(i)$ with its true expectation. Intuitively, since all relevant parameters are sharply concentrated, one would expect that the conditional expectation for any of the likely values is close to the true expectation. Making this notion precise turns out to be surprisingly cumbersome.

LEMMA 13. *There is a constant $\tau > 0$ such that, for any $q \in I^q$ and $\mathbf{b} \in I^b$, we have*

$$|\mathrm{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathrm{E}[B_j(i)]| = O(n^{1-\tau}) \ .$$

Proof. We first compare the conditional expectations for two "scenarios" of queue lengths and untouched vertices when the scenarios are close. We will see that the conditional expectations in those two scenarios will be close; from that, we can then conclude that any conditional expectation is close to the true expectation.

Given q, q' and $\mathbf{b}, \mathbf{b}'$, such that $|q - q'| \leq 4n^{1-\beta}$, and $\left|b_j - b_j'\right| \leq 2n^{1/2+\epsilon}$ for each j, we let $\hat{q} = \min(q, q')$ and $\hat{b}_j = \min(b_j, b_j')$, and define the events $\mathcal{E} := \mathcal{E}_{q,\mathbf{b}}$, $\mathcal{E}' := \mathcal{E}_{q',\mathbf{b}'}$, and $\hat{\mathcal{E}} := \mathcal{E}_{\hat{q},\hat{\mathbf{b}}}$. Now, we claim that, for some $\tau > 0$,

$$\left|\mathrm{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}] - \mathrm{E}\left[V_{\text{unto},j}(t_i) \mid \hat{\mathcal{E}}\right]\right| = O(n^{1-\tau})$$

for all j, and similarly for $\mathcal{E}'$. By the triangle inequality, this immediately implies that

$$\left|\mathrm{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}] - \mathrm{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}']\right| = O(n^{1-\tau}) \ .$$

To prove the claim, imagine that in the $(q, \mathbf{b})$ instance, we color an arbitrary, but fixed, set of $q - \hat{q}$ of copies in the queue black, as well as the copies of an arbitrary set of $b_j - \hat{b}_j$ vertices for each degree j. To expose the matching, we first expose all the neighbors of black copies, and color them blue, and then choose a uniform matching among the remaining (white, say) copies. The number of blue copies obeys some distribution $D_{q, \mathbf{b}}$, but in any case, it never exceeds the total number of black copies. Since $q, q' \in I^q$ and $\mathbf{b}, \mathbf{b}' \in I^b$, for any $\nu > 0$, this total number is at most

$$(q - \hat{q}) + \sum_j j \cdot (b_j - \hat{b}_j)$$
$$\leq 4n^{1-\beta} + 2n^{1/2+\epsilon} \sum_{j \leq n^\nu} j + 2 \sum_{j > n^\nu} j \cdot a_j n$$
$$= 4n^{1-\beta} + O(n^{1/2+\epsilon+2\nu}) + O(n^{1-(\alpha-2)\nu})$$
$$= O(n^{1-\tau}) \ ,$$

for any $\tau < \min(\beta, 1/2 - \epsilon - 2\nu, (\alpha - 2)\nu)$. Note that $\tau > 0$ as long as $1/2 - \epsilon - 2\nu > 0$; recall that we took $\epsilon < 1/2$ in the previous lemma, so we can choose any $\nu < (1/2 - \epsilon)/2$.

Now, in the $(\hat{q}, \hat{\mathbf{b}})$ instance, we can generate a uniformly random matching as follows: we choose a number k according to $D_{q, \mathbf{b}}$, choose k copies uniformly at random and color them blue, and determine a uniformly random matching among the blue copies only. Then, we match up the remaining white copies uniformly at random. We will call a node black if at least one of its copies is black, blue if at least one of its copies is blue, and white otherwise.

Since the set of nodes that are not black is deterministically the same in both instances, and the probability distribution of blue nodes is the same in both, the expected number of white nodes that end up with visible degree j is the same in both experiments. Hence, the expected total number of nodes with observed degree j can only differ by the number of blue or black nodes. Even if the degrees of those nodes were chosen adversarially, the difference cannot be more than $O(n^{1-\tau})$, since this is a deterministic upper bound on the number of black or blue copies, and hence on the number of black or blue nodes. By summing up over the entire probability space, this now proves the claim for $\mathcal{E}$ and $\hat{\mathcal{E}}$, and thus also for $\mathcal{E}$ and $\mathcal{E}'$.

We know that if $q, q' \in I^q$ and $\mathbf{b}, \mathbf{b}' \in I^b$, then they always satisfy the necessary conditions, and hence the conditional expectations are within $O(n^{1-\tau})$. Summing up over all $q \in I^q$ and $\mathbf{b} \in I^b$ therefore shows that

$$|\mathrm{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathrm{E}[B_j(i) \mid \mathcal{E}_\leq]| = O(n^{1-\tau}) \ .$$

Finally, because $\mathcal{E}_\leq$ has overwhelmingly high probability, and $B_j(i)$ is bounded by n, we obtain that, for all c,

$$|\mathrm{E}[B_j(i) \mid \mathcal{E}_\leq] - \mathrm{E}[B_j(i)]| = O(n^{-c}) \ ,$$

and the triangle inequality completes the proof. ∎

5. GENERATING FUNCTIONS

In this section, we use the formalism of generating functions [25] to express the results of Section 3 more succinctly, and complete the proof of Theorem 2. Given the generating function of the degree sequence of the underlying graph

$$g(z) = \sum_i a_i z^i \ ,$$

our goal is to obtain the generating function for the expected degree sequence of the breadth-first tree as approximated by Lemma 9,

$$g^{\mathrm{obs}}(z) = \sum_i a_i^{\mathrm{obs}} z^i \ .$$

Using the generating function formalism, we can write

$$c_{\mathrm{unto}}(t) = tg'(t), \ \delta = g'(1), \ c_{\mathrm{unex}}(t) = t^2 g'(1) \ ,$$

and from (7) we have

$$p_{\mathrm{vis}}(t) = \sum_k \frac{ka_k t^k}{tg'(t)} \left(\frac{g'(t)}{tg'(1)}\right)^k$$
$$= \frac{1}{tg'(t)} \sum_k ka_k \left(\frac{g'(t)}{g'(1)}\right)^k$$
$$= \frac{1}{tg'(1)} g'\left(\frac{g'(t)}{g'(1)}\right) \ . \tag{10}$$

Then, combining (7) and (8), the generating function for the observed degree sequence is given by

$$g^{\mathrm{obs}}(z) = \sum_m a_{m+1}^{\mathrm{obs}} z^{m+1}$$
$$= z \sum_i a_i \sum_{m=0}^{i-1} z^m \left[\int_0^1 it^{i-1} \binom{i-1}{m} \right.$$
$$\left. p_{\mathrm{vis}}(t)^m (1 - p_{\mathrm{vis}}(t))^{i-1-m} \, dt\right]$$
$$= z \sum_i a_i \left[\int_0^1 it^{i-1} \sum_{m=0}^{i-1} \binom{i-1}{m} \right.$$
$$\left. (zp_{\mathrm{vis}}(t))^m (1 - p_{\mathrm{vis}}(t))^{i-1-m} \, dt\right]$$
$$= z \sum_i a_i \int_0^1 it^{i-1} (1 - (1-z)p_{\mathrm{vis}}(t))^{i-1} \, dt$$
$$= z \int_0^1 \sum_i a_i i \cdot [t(1 - (1-z)p_{\mathrm{vis}}(t))]^{i-1} \, dt$$
$$= z \int_0^1 g'[t(1 - (1-z)p_{\mathrm{vis}}(t))] \, dt$$
$$= z \int_0^1 g'\left[t - \frac{1-z}{g'(1)} g'\left(\frac{g'(t)}{g'(1)}\right)\right] \, dt \ .$$

which completes the proof of Theorem 2.

Our definition of "reasonable" degree sequences implies that the graph is w.h.p. connected, so that every copy is eventually added to the queue. For other degree sequences, Molloy and Reed [18, 19] established that w.h.p. there is a unique giant component if $\sum_j a_j(j^2 - 2j) > 0$, and calculated its size within $o(n)$. We omit the details, but $g^{\mathrm{obs}}(z)$ is then given by an integral from t_0 to 1, where t_0 is the time at which the giant component has w.h.p. been completely exposed; this is the time at which $c_{\mathrm{unto}}(t) = c_{\mathrm{unex}}(t)$, namely the largest root less than 1 of the equation

$$\sum_j ja_j t^j = t^2 \sum_j ja_j \ . \tag{11}$$

6. EXAMPLES

6.1 Regular graphs

Random regular graphs present a particularly attractive application of the machinery developed here, as the generating function for a δ-regular degree sequence is simply $g(z) = z^\delta$. From (1), we derive the generating function for the observed degree sequence:

$$g^{\text{obs}}(z) = z\delta \cdot \int_0^1 t^{\delta-1}(1-(1-z)t^{\delta(\delta-2)})^{\delta-1}\mathrm{d}t \ . \quad (12)$$

This integral can be expressed in terms of the hypergeometric function ${}_2F_1$ [23]. In general, for all $a > -1$ and $b > 0$, we have

$$\int_0^1 t^a(1-xt^b)^{-c}\,\mathrm{d}t = \frac{1}{a+1}\,{}_2F_1\left(\frac{a+1}{b},\ c;\ \frac{a+b+1}{b};\ x\right) \ .$$

where

$$_2F_1(s,t;u;z) = \sum_{i=0}^\infty \frac{\Gamma(s+i)}{\Gamma(s)}\frac{\Gamma(t+i)}{\Gamma(t)}\frac{\Gamma(u)}{\Gamma(u+i)}\frac{z^i}{i!} \ .$$

In (12), $a = \delta - 1$, $b = \delta(\delta-2)$, and $c = 1-\delta$ (note $a > -1$ and $b > 0$ since $\delta > 2$) giving

$$g^{\text{obs}}(z) = z \cdot {}_2F_1\left(\frac{1}{\delta-2}, 1-\delta; 1+\frac{1}{\delta-2}; 1-z\right) \ . \quad (13)$$

Another useful identity is that for any negative integer q,

$$_2F_1(p,q;r;x)$$
$$= \frac{\Gamma(r)\,\Gamma(r-p-q)}{\Gamma(r-p)\,\Gamma(r-q)}\,{}_2F_1(p,q;p+q+1-r;1-x) \ .$$

Here, $q = c = 1-\delta$, and δ is an integer greater than 2. Thus, (13) becomes

$$g^{\text{obs}}(z) = z \cdot \frac{\Gamma(1+\frac{1}{\delta-2})\,\Gamma(\delta)}{\Gamma(\delta+\frac{1}{\delta-2})} \cdot {}_2F_1\left(\frac{1}{\delta-2}, 1-\delta; 1-\delta; z\right)$$

$$= z \cdot \frac{\Gamma(\delta)}{\Gamma(\delta+\frac{1}{\delta-2})\,(\delta-2)} \sum_{m=0}^\infty \Gamma\left(m+\frac{1}{\delta-2}\right)\frac{z^m}{m!} \ .$$

The expected observed degree sequence is then given by

$$a_{m+1}^{\text{obs}} = \frac{\Gamma(\delta)\,\Gamma(m+\frac{1}{\delta-2})}{\Gamma(\delta+\frac{1}{\delta-2})\,(\delta-2)\,m!} \ .$$

To explore the asymptotic behavior of a_{m+1}^{obs}, note that

$$\Gamma(m) < \Gamma(m+\epsilon) < \Gamma(m)\,m^\epsilon$$

for all $m \geq 2$ and all $0 < \epsilon < 1$. Therefore, for $m \geq 2$, we can bound a_{m+1}^{obs} as follows:

$$\frac{m^{-1}}{\delta^{1/(\delta-2)}\,(\delta-2)} \ < \ a_{m+1}^{\text{obs}} \ < \ \frac{m^{-1+1/(\delta-2)}}{\delta-2} \ .$$

For any fixed δ, this gives a power-law degree sequence, and in the limit of large δ, one observes $a_{m+1}^{\text{obs}} \sim m^{-1}$. Thus, even regular graphs appear to have a power-law degree distribution (with exponent $\alpha \to 1$ in the limit $\delta \to \infty$) under traceroute sampling!

6.2 Poisson degree distributions

Clauset and Moore [6] used the method of differential equations to show that a breadth-first tree in the giant component of $G(n, p = \delta/n)$ has a power-law degree distribution, $a_{m+1} \sim m^{-1}$ for $m \lesssim \delta$. Here, we recover this result as a special case of our analysis. Recall that w.h.p. the degree distribution of $G(n, p = \delta/n)$ is within $o(1)$ of a Poisson degree sequence with mean δ. The generating function is then $g(z) = e^{-\delta(1-z)}$, and the generating function for the observed degree sequence is

$$g^{\text{obs}}(z) = z\delta \cdot \int_{t_0}^1 e^{-\delta(1-t)}e^{-\delta(1-z)e^{-\delta(1-e^{-\delta(1-t)})}}\mathrm{d}t$$

$$= z \int_0^{1-t_0} e^{-\delta(1-z)e^{-\delta y}}\mathrm{d}y \ . \quad (14)$$

In the second integral we transform variables by taking $y = 1 - e^{-\delta(1-t)}$. Here, t_0 is the time at which we have exposed the giant component, i.e., when $c_{\text{unex}}(t) = c_{\text{unto}}(t)$; since $c_{\text{unex}}(t) = \delta t^2$ and $c_{\text{unex}}(t) = \delta t e^{-\delta(1-t)}$, t_0 is the smallest positive root of $t = e^{-\delta(1-t)}$.

This integral can be expressed in terms of the exponential integral function $\text{Ei}(z)$ [24] and the incomplete Gamma function $\Gamma(a,z)$, which are defined as

$$\text{Ei}(z) = -\int_{-z}^\infty \frac{e^{-x}}{x}\,\mathrm{d}x$$

$$\Gamma(a,z) = \int_z^\infty x^{a-1}e^{-x}\,\mathrm{d}x \ .$$

Then, with the integral

$$\int_p^q e^{ae^{by}}\mathrm{d}y = \frac{1}{b}\left(\text{Ei}\left[ae^{qb}\right] - \text{Ei}\left[ae^{pb}\right]\right)$$

and the Taylor series

$$\text{Ei}(-\delta(1-z)) = \text{Ei}(-\delta) - \sum_{k=1}^\infty \frac{\Gamma(k,\delta)}{\Gamma(k)}\frac{z^k}{k} \ ,$$

taking $a = -\delta(1-z)$ and $b = -\delta$ as in (14) gives

$$g^{\text{obs}}(z) \approx \frac{z}{\delta}\left(\text{Ei}\left[-\delta(1-z)\right] - \text{Ei}\left[-\delta e^{-\delta(1-t_0)}(1-z)\right]\right)$$

$$= \sum_{m=0}^\infty \frac{z^{m+1}}{\delta m!}\left(\Gamma(m, \delta e^{-\delta(1-t_0)}) - \Gamma(m,\delta)\right) \ . \quad (15)$$

Thus, the coefficients of the observed degree sequence are

$$a_{m+1}^{\text{obs}} = \frac{1}{\delta m!}\int_{\delta e^{-\delta(1-t_0)}}^\delta e^{-x}x^{m-1}\,\mathrm{d}x \ . \quad (16)$$

Now, t_0 approaches $e^{-\delta}$ in the limit of large δ, and for $m \lesssim \delta$, the integral of (16) coincides almost exactly with the full Gamma function $\Gamma(m)$ since it contains the peak of the integrand. Specifically, in [6] Clauset and Moore showed that if $m < \delta - \delta^\kappa$ for some $\kappa > 1/2$, then

$$a_{m+1}^{\text{obs}} = (1-o(1))\frac{\Gamma(m)}{\delta m!} \ \sim \ \frac{1}{\delta m} \ ,$$

giving an observed degree sequence of power-law form m^{-1} up to $m \sim \delta$ and confirming the experimental result of Lakhina et al. [14].

7. CONCLUSIONS

Having established rigorously that single-source traceroute sampling is biased, thus formally verifying the empirical observations of Lahkina et al. [14], and having calculated the precise nature of that bias for a broad class of random graphs, there are several natural questions we may now ask.

Petermann and De Los Rios [22] and Clauset and Moore [6] both demonstrated experimentally that when the graph does have a power-law degree distribution, traceroute sampling can significantly underestimate the exponent α. Although a characterization of this phenomenon is beyond the scope of this paper, it is a natural application of our machinery.

However, a more intriguing question is the following: can we invert Theorem 2, and derive $g(z)$ from $g^{\mathrm{obs}}(z)$? In other words, can we undo the bias of traceroute sampling, and infer the most likely underlying distribution given the observed distribution? Unfortunately, it is not even clear whether the mapping from $g(z)$ to $g^{\mathrm{obs}}(z)$ is invertible, and the complexity of our expression for $g^{\mathrm{obs}}(z)$ makes such an inversion appear quite difficult. We leave this question for future work.

Finally, although several studies claim that using additional sources in mapping the Internet has only a small marginal utility [2, 21], Clauset and Moore [6] showed empirically that in power-law random graphs, the number of sources required to compensate for the bias in traceroute sampling grows linearly with the mean degree of the network. However, a rigorous analysis of multiple sources seems quite difficult, since the events that a given edge appears in BFS trees with different roots are highly correlated. We leave the generalization of our results to traceroute sampling with multiple sources for future work.

Acknowledgments

8. REFERENCES

[1] L. Amini, A. Shaikh, and H. Schulzrinne. Issues with inferring Internet topological attributes. In *SPIE IT-Com*, 2002.

[2] P. Barford, A. Bestavros, J. Byers, and M. Crovella. On the marginal utility of network topology measurements. In *SIGCOMM Internet Measurement Workshop*, 2001.

[3] B. Bollobás. *Random graphs*. Academic Press, 1985.

[4] B. Bollobás and F. Chung. The diameter of a cycle plus a random matching. *SIAM J. Disc. Math.*, 1:328–333, 1988.

[5] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. In *Proc. 21st ACM SIGCOMM Conference*, 2002.

[6] A. Clauset and C. Moore. Accuracy and scaling phenomena in Internet mapping. *Physical Review Letters*, 94:018701, 2005.

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

[8] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the Internet. In *Proc. 29th Intl. Colloq. on Automata, Languages and Programming*, pages 110–122, 2002.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. ACM SIGCOMM Conference*, pages 251–262, 1999.

[10] R. Govindan and H. Tangmunarunkit. Heuristics for Internet map discovery. In *Proc. 19th ACM SIGCOMM Conference*, 2000.

[11] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. of the American Statistical Association*, 58:13–30, 1963.

[12] J. H. Kim. Poisson cloning model for random graphs. Preprint, 2004.

[13] J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. In *Intl. Conf. on Combinatorics and Computing*, 1999.

[14] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *Proc. 22nd ACM SIGCOMM Conference*, 2003.

[15] J. Leguay, M. Latapy, T. Friedman, and K. Salamatian. Describing and simulating internet routes. Preprint `cs.NI/0411051`, 2004.

[16] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–247. Springer, 1998.

[17] M. Mihail, C. Papadimitriou, and A. Saberi. On certain connectivity properties of the Internet topology. In *Proc. 35th ACM Symp. on Theory of Computing*, 2003.

[18] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–180, 1995.

[19] M. Molloy and B. Reed. The size of the largest component of a random graph on a fixed degree sequence. *Combinatorics, Probability and Computing*, 7:295–306, 1998.

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

[21] J.-J. Pansiot and D. Grad. On routers and multicast trees in the Internet. *ACM SIGCOMM Communication Review*, 28:41–50, 1998.

[22] T. Petermann and P. De Los Rios. Exploration of scale-free networks: do we measure the real exponents? *Euro. Phys. Journal B*, 38:201, 2004.

[23] J. B. Seaborn. *Hypergeometric Functions and Their Applications*. Springer, 1991.

[24] J. Spanier and K.B. Oldham. The exponential integral ei(x) and related functions. In *An Atlas of Functions*, chapter 37, pages 351–360. Hemisphere, 1987.

[25] H.S. Wilf. *generatingfunctionology*. Academic Press, 1994.

[26] N.C. Wormald. Models of random regular graphs. In J. Lamb and D. Preece, editors, *London Math. Soc. Lecture Note Series*, volume 276, pages 239–298. Cambridge University Press, 1999.