On the Bias of Traceroute Sampling: or, Power-Law Degree Distributions in Regular Graphs

DIMITRIS ACHLIOPTAS

University of California, Santa Cruz

AARON CLAUSET

University of New Mexico, Albuquerque, and the Santa Fe Institute, New Mexico

DAVID KEMPE

University of Southern California, Los Angeles

AND

CRISTOPHER MOORE

University of New Mexico, Albuquerque, and the Santa Fe Institute, New Mexico

Abstract. Understanding the graph structure of the Internet is a crucial step for building accurate network models and designing efficient algorithms for Internet applications. Yet, obtaining this graph structure can be a surprisingly difficult task, as edges cannot be explicitly queried. For instance, empirical studies of the network of Internet Protocol (IP) addresses typically rely on indirect methods like *traceroute* to build what are approximately single-source, all-destinations, shortest-path trees. These trees only sample a fraction of the network's edges, and a paper by Lakhina et al. [2003] found empirically that the resulting sample is intrinsically biased. Further, in simulations, they observed

A preliminary version of this article appeared in *Proceedings of the ACM Symposium on Theory of Computing*, ACM, New York, 2005, pp. 674–683.
The work of D. Achlioptas was done, in part, while with Microsoft Research, and was supported by NSF CAREER award CCF-0546900, an Alfred P. Sloan Fellowship, and ERC IDEAS grant 210743. The work of A. Clauset was supported by NSF grant ITR-0324845 and PHY-0200909. The work of D. Kempe was supported by an NSF Postdoctoral Fellowship. The work of C. Moore was supported by NSF grants CCF-0524613, CCR-0220070, EIA-0218563, ITR-0324845, and PHY-0200909.
Authors' addresses: D. Achlioptas, Department of Computer Science, University of California Santa Cruz, Santa Cruz, CA 95064, e-mail: optas@cs.ucsc.edu; A. Clauset, 1399 Hyde Park Road, Santa Fe, NM 87501, e-mail: aaronc@santafe.edu; D. Kempe, Department of Computer Science, University of Southern California, Los Angeles, CA 90089, e-mail: dkempe@usc.edu; C. Moore, Department of Computer Science, 1 University of New Mexico, MSC01 1130, Albuquerque, NM 87131, e-mail: moore@cs.unm.edu.

DOI 10.1145/1538902.1538905 http://doi.acm.org/10.1145/1538902.1538905

that the degree distribution under traceroute sampling exhibits a power law even when the underlying degree distribution is Poisson.

In this article, we study the bias of traceroute sampling mathematically and, for a very general class of underlying degree distributions, explicitly calculate the distribution that will be observed. As example applications of our machinery, we prove that 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: C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*; C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*; G.3 [**Probability and Statistics**]: *Stochastic processes*

General Terms: Measurement, Reliability, Theory

Additional Key Words and Phrases: Internet topology, traceroute, sampling bias

ACM Reference Format:

Achlioptas, D., Clauset, A., Kempe, D., and Moore, C. 2009. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. J. ACM 56, 4, Article 21 (June 2009), 28 pages. DOI = 10.1145/1538902.1538905 http://doi.acm.org/10.1145/1538902.1538905

1. *Introduction*

A large body of recent work has focused on the topological properties of the Internet. Perhaps most famously, Faloutsos et al. [1999] claimed a power-law degree distribution both in the graph of routers in the Internet, that is, the level at which the Internet Protocol (IP) operates, and the connections between autonomous systems, the level at which the Border Gateway Protocol (BGP) operates. Similar results were obtained by Govindan and Tangmunarunkit [2000] and Barford et al. [2001], among others. Based on these and other topological studies, it is widely believed that the degree distribution of Internet routers has a power-law form with exponent $2 < \alpha < 3$, that is, 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, Fabrikant et al. [2002]) and research into the question of how the topology might affect the real-world performance of Internet algorithms and mechanisms (for instance, Mihail et al. [2003] and Alderson et al. [2005]).

However, unlike graphs such as the World Wide Web [Kleinberg et al. 1999], in which links from each site can be readily observed, the connections between IP-level routers on the Internet cannot be queried directly. Without explicitly knowing which routers are connected, how can one obtain an accurate map of this level 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 IP-level Internet. While actual routing decisions involve multiple protocols and network layers, it is a common and reasonably accurate assumption that the packets follow shortest paths between their source and destination. We discuss the validity of this assumption and its implications in more detail below.

Most studies, including the one on which the work of Faloutsos et al. [1999] is based, infer the Internet's IP-level topology from the union of traceroutes issued from a single root computer to a large number of other computers in the network. If each edge has unit cost plus a small random term, the union of these shortest paths is

a breadth-first search (BFS) tree.[1] This model of the sampling process is admittedly an idealization for several reasons. First of all, most empirical studies only use a subset of the valid IP addresses as destinations. Second, for technical reasons, some routers may not respond to traceroute queries. Third, a single router may annotate different packets with different IP addresses, a problem known as *interface aliasing*. Similarly, several routers may share a single IP address by communicating with each other using a data link layer protocol, as in a token ring network. These issues are known to introduce noise into the measured topology [Amini et al. 2002; Chen et al. 2002].

However, as Lakhina et al. [2003] recently pointed out, traceroute sampling has another fundamental bias, one that 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 properties being sampled, such as node degrees. However, an edge is much more likely to be *visible*, that is, included in the BFS tree or the traceroute study, if it is close to the root of the tree. 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, Lakhina et al. [2003] (and, more recently, Guillaume et al. [2006]) showed empirically that for Erdős-Rényi random graphs $G(n, p)$ [Erdős and Rényi 1960], which have a Poisson degree distribution, the observed degree distribution under traceroute sampling follows a power law, and this was verified analytically by Clauset and Moore [2005]. In other words, the bias introduced by traceroute sampling alone can make power laws appear where none exist in the underlying graph. Even when the underlying graph actually does have a power-law degree distribution $k^{-\alpha}$, Petermann and De Los Rios [2004] and Clauset and Moore [2005] showed numerically that traceroute sampling can significantly underestimate its exponent α, especially when the average degree of the underlying graph is large.

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

Our answers to these questions demonstrate that formal mathematical analysis can highlight the inherent problems with the current methodologies used to map the structure of the IP-level Internet. That is, we quantify precisely the bias introduced by traceroute-based sampling, assuming that traceroutes are shortest paths, while verifying formally the empirical observations of Lakhina et al. [2003] and Guillaume et al. [2006]. In particular, we prove that traceroute sampling reports power-law degree distributions even in the extreme case of *regular* random graphs, where every vertex has the same degree! Moreover, our results can be viewed as a first step towards the more practical goal of inferring a graph's true underlying degree distribution from biased observations.

[1] Several studies, including Barford et al. [2001] and Pansiot and Grad [1998], have combined traceroutes from multiple sources. However, the number of sources used is typically quite small (about 12). The NetDimes project [Shavitt and Shir 2005] uses many more sources, but samples many fewer destinations.

1.1. OUR RESULTS. Our main result in this article is Theorem 1.2, which gives an explicit, high-probability characterization of the observed degree distribution as a function of the true underlying distribution. 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.

Definition 1.1. A degree distribution $\{a_k\}$ is *sober* 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.

We restrict our attention to distributions that are *sober* as defined above, that is, distributions that are "not too heavy-tailed," and in which all nodes have degree at least 3. The requirement that the minimum degree is at least 3 implies, through a simple counting argument, that the graph is with high probability 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. 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 $\alpha \leq 3$). In particular, this requirement is consistent with the conjectured range $\alpha \in (2, 3)$ for the Internet [Faloutsos et al. 1999; Govindan and Tangmunarunkit 2000].

In order to speak precisely about a random (multi)graph with a given degree sequence, we will use the *configuration model* [Bollobás 2001]: 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. Standard estimates imply that if a degree sequence is sober, then with high probability the total number of self-loops and multiple edges is $o(n)$, and no vertex has more than $C \log n$ such edges, for some constant C. Thus, the difference between such a graph and a random simple graph is negligible, and we may work with the multigraph as a matter of mathematical convenience. Our main result can then be stated as follows:

THEOREM 1.2. *Let $\{a_j\}$ be a sober 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. There exists a constant $\zeta > 0$ such that with high probability, $|A_j^{\text{obs}} - a_j^{\text{obs}} n| < n^{1-\zeta}$ for all j, where*

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

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

[2] We say that a sequence of events $\mathcal{E}_n$ occurs *with high probability* if $\text{Prob}[\mathcal{E}_n] = 1 - o(1)$ as $n \to \infty$, and *with overwhelmingly high probability* if $\text{Prob}[\mathcal{E}_n] = 1 - o(n^{-c})$ 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.

More concisely, if $g(z) = \sum_{j=0}^{\infty} a_j z^j$ *is the generating function of the degree sequence, then* a_j^{obs} *is the coefficient of* z^j *in*

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

The bulk of this article, namely Sections 2–5, is devoted to the proof of Theorem 1.2. In Section 6, we apply our general result to δ-regular graphs and graphs with Poisson degree distributions, which serve as conceptually crisp examples of the extreme bias that traceroute sampling can exhibit. 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 [2005].

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 somewhat analogous to Kim's Poisson cloning model [Kim 2006], 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 the degree of a vertex 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 1.2. The concentration part of the result, in Section 4, analyzes a different and much more coarse-grained view of the process using tools distinct from those in Section 3: by carefully conditioning on the history of the process, we can apply a small number of Martingale-style bounds to obtain overall concentration.

1.2. Caveats Regarding Traceroute. As discussed above, we implicitly assume that the paths discovered by traceroute are shortest paths. This assumption is only an approximation for several reasons. First, as we discussed above, the mapping between routers and IP addresses (which are what is observed in reality) is neither one-to-one nor well-defined. This creates aliasing problems, among others [Spring et al. 2004]. Traceroute itself can also introduce spurious edges: traceroute works by sending a series of packets toward the destination, with increasing bounds on the number of hops, and seeing where each of these packets terminates. Since congestion and load balancing can cause these packets to occasionally take different routes, paths of k and $k+1$ hops can have nonadjacent endpoints, in which case traceroute will incorrectly infer an edge between them. These and other traceroute artifacts are discussed in more detail by Viger et al. [2006] who propose an improved version of traceroute, called "Paris traceroute," which significantly reduces some, but not all, of these issues.

While the above are measuring artifacts, a more fundamental problem is introduced by interactions between routing protocols at different layers of the TCP/IP framework. For instance, actual routes are strongly affected by routing policies implemented by autonomous systems (ASs) at the BGP level; these are often based in part on contractual obligations or business concerns, in addition to latency (path length) concerns. Also, a shortest path at the AS level is not necessarily equivalent to a shortest path at the router level [Tangmunarunkit et al. 2001], or vice-versa. Thus, some routes can be up to several hops longer than the actual shortest paths [Tangmunarunkit et al. 2001; Leguay et al. 2007]. On the other hand, such

issues do not fundamentally change the manner in which traceroute reveals edges in the IP-level Internet, suggesting that the idealizations made by our model may be reasonable. Similar conclusions have been reached by Guillaume et al. [2006], through extensive simulations.

More broadly, our choice of model and analysis is motivated mainly by the desire to understand the fundamental bias introduced by the BFS tree sampling technique alone, rather than the artifacts of a particular software tool or routing policy. As IP-level sampling tools improve (e.g., Viger et al. [2006]) and access to BGP routing tables (which are typically used to reconstruct the AS-level Internet graph) becomes more widely available, one might reasonably expect the idealized analysis to gradually mirror the real world more accurately. In addition, our results are not necessarily specific to traceroute sampling, but applicable to the exploration of other types of networks by probing shortest paths from a single source.

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, that is, it still has 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 become *visible*. On the other hand, edges to neighbors that are already in the queue are not visible.

As we will carry out the analysis in the configuration model, we need to describe 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. In describing the exploration process, we refer to two copies of the same vertex as *siblings*. Each copy will be either *exposed*, *enqueued*, or *untouched*. An exposed copy is one whose partner has already been revealed. An unexposed copy is enqueued if it is in $\mathcal{Q}$, and *untouched* otherwise. At each step, the copy u at the head of the queue has its partner v *exposed*, and both of them are removed from the matching. At this point, all of v's siblings are added to the queue, unless they were in the queue already. The possible states of copies relate to the states of the corresponding vertices as described in Algorithm 2.1. 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 search is as described in Algorithm 2.1.

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 [Motwani and Raghavan 1990], 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

Algorithm 2.1: Breadth-First Search at the Copy Level

```
1:  while Q is nonempty do
2:      Pop a copy u from the head of 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 Q
7:      else
8:          Remove v from Q
9:      end if
10: end while
```

[0, 1]. Then, at each step, u's partner v is chosen as the unexposed copy with the the largest index. Thus, we can 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.

This continuous-time description is really just a particular way to generate a uniformly random matching and explore it at the same time. However, it makes certain quantities independent that would otherwise be correlated in complicated ways, and is thus very convenient for our analysis. In particular, since the indices of v's siblings are uniformly random conditional on being less than t, this description 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 $C_{\text{unex}}(t)$ and $C_{\text{unto}}(t)$ denote the number of unexposed and untouched copies, and let $V_{\text{unto},j}(t)$ denote the number of untouched vertices of degree j; note that $C_{\text{unto}}(t) = \sum_j jV_{\text{unto},j}(t)$. 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, $\mathrm{E}[V_{\text{unto},j}(t)] = a_j t^j n$, and

$$\mathrm{E}[C_{\text{unto}}(t)] \;=\; \textstyle\sum_j ja_j t^j n \;=:\; c_{\text{unto}}(t)\cdot n. \tag{2}$$

To calculate $\mathrm{E}[C_{\text{unex}}(t)]$, 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

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

The following lemma shows that $V_{\text{unto},j}(t)$, $C_{\text{unto}}(t)$ and $C_{\text{unex}}(t)$ 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 2.1. *Let $\{a_j\}$ be a sober 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 $j < n$, with overwhelmingly high probability:*

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

where $c_{\text{unto}}(t)$ and $c_{\text{unex}}(t)$ are given by Eqs. (2) *and* (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 [Hoeffding 1963] (Theorem 3 from McDiarmid [1998]):

THEOREM 2.2. *If $X_1, \ldots, X_k$ are independent, non-negative random variables with $X_i \le b_i$ for all i, and $X = \sum_i X_i$, then for any $\Delta \ge 0$:*

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

First, $V_{\text{unto},j}(t)$ is a binomial random variable distributed as $\text{Bin}(a_j n, t^j)$. By applying Theorem 2.2 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 $2\text{e}^{-2n^{2\epsilon}/a_j} \le \text{e}^{-n^{2\epsilon}}$. Thus, at each individual time t and for each j, the stated bound on $V_{\text{unto},j}(t)$ holds with overwhelmingly high probability.

We wish to show that this bound holds with overwhelmingly high probability for all j and all t, that is, 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 sober). 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

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

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 2.2

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 2.2 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, whenever $\beta < \min(\frac{1}{2}, \frac{\alpha-2}{2})$, we obtain $|C_{\text{unto}}(t) - \mathrm{E}[C_{\text{unto}}(t)]| \le n^{1-\beta}$ with overwhelmingly high probability, by Theorem 2.2. A union bound over t as before completes the proof. □

3. *Expected Degree Distribution*

In this section, we begin the proof of Theorem 1.2 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, that is, that a vertex of degree i has m copies that give rise to visible edges. Then, the expected degree distribution is given by

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

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 $\mathrm{d}t^i/\mathrm{d}t = it^{i-1}$, and we have

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

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. As defined in Algorithm 2.1, 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.

The probability of this event is much easier to calculate if we make several independence assumptions. In doing so, we only calculate an approximation to the quantity we are interested in. Below, in a series of lemmas, we will show that the independence assumptions and approximations hold with high probability for all but a negligible fraction of the vertices. We can thus with high probability bound the deviation of the actual degree distribution from the approximation we calculate next. As a result, the following analysis is completely rigorous, despite our initial approximations.

For now, we assume that untouched copies are chosen without replacement, and that the untouched copy taken for w is still available. In addition, we assume that v, its neighbors, and its neighbors' neighbors form a tree, that is, that v does not occur in a triangle or 4-cycle, and that neither it nor its neighbors have any parallel edges. Then, the probability that the partners of w's $k-1$ siblings are all untouched is $P_{\text{unto}}(t)^{k-1}$. This gives

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

Since $V_{\text{unto},k}(t)$, $C_{\text{unto}}(t)$ and $C_{\text{unex}}(t)$ are concentrated, per Lemma 2.3, substituting their expectations instead of the variable itself gives only lower-order errors with high probability, and these errors are absorbed into the overall error bounds. We thus obtain the following approximation for $P_{\text{vis}}(t)$:

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

Approximating further, 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). Then, 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 $\text{Bin}(i-1, p_{\text{vis}}(t))$. This analysis then gives us the degree distribution stated in Theorem 1.2. However, it remains to be shown that the multiple approximations made in the analysis only lead to lower-order error terms.

To this end, we wish to confirm the above heuristic analysis by showing that with high probability 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 with high probability 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 2.3.

LEMMA 3.1. *The probability that a random copy belongs to a vertex of degree greater than k is* $o(k^{-2\beta})$.

PROOF. This probability is

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

LEMMA 3.2. *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 3.1 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, that is, 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, that is, 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 3.2, 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 3.3. *Let η, γ be defined as in Lemma* 3.2. *There exists a constant $\theta > 0$ such that for $t \in [n^{-\theta}, 1]$ and $i < n^{-\eta}$,*

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

PROOF. Let $d_{\min}$ be the minimum degree of the graph, that is, 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 with overwhelmingly high probability by Lemma 2.3.

Conditioning on v's neighbors having degree at most n^{λ} as in Lemma 3.2, 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 3.2 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 Eq. (6). □

The next three lemmas then show that $p_{\text{vis}}(t)$ is a very good approximation for $P_{\text{vis}}(t)$ for most t, and that therefore the distribution of $\rho_{i,m}(t)$ is very close to $\text{Bin}(i-1, p_{\text{vis}}(t))$.

LEMMA 3.4. *Let γ and θ be defined as in Lemma* 3.2 *and Lemma* 3.3. *There exists a constant $\kappa > 0$ such that for all $t \in [n^{-\theta}, 1 - n^{-\kappa}]$, with overwhelmingly high probability $|P_{\text{vis}}(t) - p_{\text{vis}}(t)| < n^{-\gamma}$.*

PROOF. Recall that $\theta = \beta/(2d_{\min})$. From Lemma 2.3, since $t \geq n^{-\theta}$ we have with overwhelmingly high probability $C_{\text{unex}}(t) = \Omega(t^2 n) = \Omega(n^{1-\beta/d_{\min}})$ as in the previous lemma, and $C_{\text{unto}}(t) = \Omega(t^{d_{\min}} n) = \Omega(n^{1-\beta/2})$. For definiteness, take $\epsilon = 1/12$ in Lemma 2.3; then with overwhelmingly high probability

$$\begin{aligned} C_{\text{unex}}(t) &= c_{\text{unex}}(t)n + o(n^{7/12}) = c_{\text{unex}}(t)n \cdot (1 + o(n^{\beta/d_{\min}-5/12})) \\ C_{\text{unto}}(t) &= c_{\text{unto}}(t)n + o(n^{1-\beta}) = c_{\text{unto}}(t)n \cdot (1 + o(n^{-\beta/2})) \\ V_{\text{unto},k}(t) &= a_k t^k n + o(n^{7/12}). \end{aligned}$$

Recall that $\beta < 1/2$ and $d_{\min} \geq 3$. Since $\beta/d_{\min} - 5/12 < -1/4 < -\beta/2$,

$$P_{\text{unto}}(t) = \frac{C_{\text{unto}}(t)}{C_{\text{unex}}(t)} = \frac{c_{\text{unto}}(t)}{c_{\text{unex}}(t)}\left(1 + o(n^{-\beta/2})\right).$$

Now, we compare $P_{\text{vis}}(t)$ with $p_{\text{vis}}(t)$ term by term, and separate their respective sums into the terms with $3 \le k \le n^{\beta/12}$ and those with $k > n^{\beta/12}$. For all $k \le n^{\beta/12}$, since $\beta/12 + \beta/2 - 5/12 < -1/8 < -\beta/4$,

$$P_{\text{unto},k}(t) \;=\; \frac{kV_{\text{unto},k}(t)}{C_{\text{unto}}(t)} \;=\; \frac{ka_k t^k}{c_{\text{unto}}(t)} + o(n^{-\beta/4}),$$

and since $(1+x)^y = 1 + O(xy)$ if $xy < 1$,

$$P_{\text{unto}}(t)^k \;=\; \left(\frac{c_{\text{unto}}(t)}{c_{\text{unex}}(t)}\right)^k \left(1 + o(n^{-\beta/2})\right)^k \;=\; \left(\frac{c_{\text{unto}}(t)}{c_{\text{unex}}(t)}\right)^k \left(1 + o(n^{-5\beta/12})\right).$$

Thus, each term obeys

$$P_{\text{unto},k}(t) P_{\text{unto}}(t)^k = \frac{ka_k t^k}{c_{\text{unto}}(t)} \left(\frac{c_{\text{unto}}(t)}{c_{\text{unex}}(t)}\right)^k + o(n^{-\beta/4})$$

and the total error from the first $n^{\beta/12}$ terms is at most $n^{\beta/12} \cdot o(n^{-\beta/4}) = o(n^{-\beta/6})$.

On the other hand, if $k > n^{\beta/12}$, then for any $t \le 1 - n^{-\kappa}$ we have

$$t^k \;<\; \mathrm{e}^{-kn^{-\kappa}} \;<\; \mathrm{e}^{-n^{\beta/12-\kappa}},$$

Setting $\kappa < \beta/12$ makes this exponentially small. In that case, taking a union bound over all $n^{\beta/12} < k < n$, with overwhelmingly high probability there are no unexposed vertices of degree greater than $n^{\beta/12}$; thus $P_{\text{unto},k}(t) = 0$ and these terms of $P_{\text{vis}}(t)$ are zero. The corresponding terms of $p_{\text{vis}}(t)$ are exponentially small as well, so the total error from these terms is exponentially small. Thus, the total error is $o(n^{-\beta/6})$, and since $\gamma = \beta^2/3 < \beta/6$, this can be absorbed into the probability $o(n^{-\gamma})$ that the conditioning of Lemma 3.2 is violated. □

LEMMA 3.5. *For any s, m, p and Δ,*

$$|\text{Prob}[\text{Bin}(s, p) = m] - \text{Prob}[\text{Bin}(s, p + \Delta) = m]| \le s\Delta.$$

PROOF. It is sufficient to bound the derivative of these probabilities with respect to p as follows.

$$\begin{aligned}
\left|\frac{\partial}{\partial p}\text{Prob}[\text{Bin}(s, p) = m]\right| &= \left|\frac{\partial}{\partial p}\binom{s}{m} p^m (1-p)^{s-m}\right| \\
&= \binom{s}{m} p^m (1-p)^{s-m} \left|\frac{m}{p} - \frac{s-m}{1-p}\right| \\
&\le \binom{s}{m} p^m (1-p)^{s-m} \max\left(\frac{m}{p}, \frac{s-m}{1-p}\right) \\
&\le \max\left(\sum_{m=0}^{s} \binom{s}{m} \frac{m}{p}\, p^m (1-p)^{s-m} \;,\right. \\
&\qquad \left. \sum_{m=0}^{s} \binom{s}{m} \frac{s-m}{1-p}\, p^m (1-p)^{s-m}\right) \\
&= \max(s, s) \;=\; s.
\end{aligned}$$

□

LEMMA 3.6. *Let $\{a_j\}$ be a sober 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,*

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

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

PROOF. Given Lemma 3.4, we apply Lemma 3.5 and the triangle inequality. In this case, we have $s \le n^{\eta}$ and $\Delta < n^{-\gamma}$, so the error in $\rho_{i,m}$ is at most $n^{\eta-\gamma}$. Recalling from the proof of Lemma 3.2 that $\eta = \beta^2/6$ and $\gamma = \beta^2/3$, for sufficiently large n this is less than $n^{-\mu}$ for any $\mu < \beta^2/6$. □

Finally, combining Lemma 3.6 with Eqs. (4), (5), and (7), if

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

where, combining Eq. (7) with Eqs. (2) and (3),

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

then we have the following lemma.

LEMMA 3.7. *Let $\{a_i\}$ be a sober degree sequence and assume that G is connected. There is a constant $\zeta > 0$ such that for sufficiently large n, for all $j < n$*

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

PROOF. There are three sources of error in our estimate of $\mathrm{E}[A_j^{\text{obs}}]$ for each j. These are the error $n^{-\mu}$ in $\rho_{i,m}(t)$ given by Lemma 3.6, 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 with overwhelmingly high probability 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 $\mathrm{E}[A_j^{\text{obs}}]$. Specifically, we prove

THEOREM 4.1. *There is a constant $\rho > 0$ such that, with overwhelmingly high probability, the following holds simultaneously for all j:*

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

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 already established in Lemma 2.3.

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

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

be the expected queue size at time t. 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

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

where β is defined as in Lemma 2.1. 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 $c_{\text{unex}}(t_i) < 2n^{-\beta}$. 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 4.2 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

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

for all times i, then all of the $\mathcal{Q}(t_i)$ are disjoint. Each of these bounds holds with overwhelmingly high probability by Lemma 2.1, and by the union bound, with overwhelmingly high probability 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 $c_{\text{unex}}(0) = 0$ implies that $c_{\text{unto}}(t_r) < 2n^{-\beta}$, so by Lemma 2.1, at most $O(n^{1-\beta})$ copies remain unexposed with overwhelmingly high probability. On the other hand, when $r = \log^2 n$, we can use that the diameter of a random graph is bounded by $\log^2 n$ with probability at least $1 - O(n^{-13/27})$, a fact we establish in Section 4.1 below.[3] Even if $C_{\text{unto}}(t_r)$ were $\Omega(n)$ in the remaining case, as 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 $|V_{\text{unto},j}(t_i) - a_j t_i^j n| \leq n^{1/2+\epsilon}$ and $|Q(t_i) - q(t_i)| \leq 2n^{1-\beta}$ hold simultaneously for all i. By Lemma 2.1, $\mathcal{E}$ occurs with overwhelmingly high probability. 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, $|A_j^{\text{obs}} - \sum_{i=1}^{r} B_j(i)| = \tilde{O}(n^{1-\beta})$ with overwhelmingly high probability, which implies that $|\mathrm{E}[A_j^{\text{obs}}] - \sum_{i=1}^{r} \mathrm{E}[B_j(i)]| = \tilde{O}(n^{1-\beta})$, since this difference is deterministically bounded above by n.

By Lemma 4.2 below and a union bound over all i, there is a $\tau > 0$ such that with overwhelmingly high probability $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$ 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 with overwhelmingly high probability:

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

for any $\rho < \min(\beta, \tau)$, completing the proof of Theorem 4.1. □

[3] In order to achieve the type of high-probability guarantee we are aiming for, we require the probability for the diameter bound to be at least $1 - O(n^{-13/27})$. Unfortunately, this probability is higher than the bounds established in the previous literature, necessitating the lengthy proof in Section 4.1.

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

LEMMA 4.2. *There is a constant $\tau > 0$ such that, for any fixed i, with overwhelmingly high probability, $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$ 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 [1999, 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 4.3 (WORMALD 1999). *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 $V_{\mathrm{unto},j}(t_i) = b_j$ 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 4.3 implies that

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

holds with overwhelmingly high probability 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 4.3 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 4.4 below that the conditional expectations $\mathrm{E}\big[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}\big]$ are close to the actual expectations $\mathrm{E}\big[B_j(i)\big]$. 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 $V_{\mathrm{unto},j}(t_i) \in I_j^b$ for all j. Notice that $\mathcal{E}_{\leq}$ occurs with overwhelmingly high probability by Lemma 2.1.

Lemma 4.4 below ensures that whenever $q \in I^q$ and $\mathbf{b} \in I^b$, then the conditional expectation is close to the true expectation, that is, for some $\tau > 0$,

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

Thus, for all such q and $\mathbf{b}$, combining this with Eq. (9) and the triangle inequality gives $|B_j(i) - \mathrm{E}[B_j(i)]| = O(n^{1-\tau})$, so the latter occurs with overwhelmingly high probability, 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 technical.

LEMMA 4.4. *There is a constant $\tau > 0$ such that, for any $q \in I^q$ and $\mathbf{b} \in I^b$, we have*

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

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

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

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

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

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

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

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

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

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

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

and the triangle inequality completes the proof. □

4.1. A High Probability Bound for the Diameter. Here we bound the diameter of a random (multi)graph with a given degree sequence. Our result is less precise than, say, that of Bollobás and Chung [1988] for random 3-regular multigraphs, but holds with higher probability, a necessity for our application.

Theorem 4.5. *Let $\{a_i\}_i$ be a degree sequence with $a_i = 0$ for $i < 3$. Let G be a random multi-graph with n nodes in the configuration model, with degree distribution $\{a_i\}$, that is, $a_i \cdot n$ nodes have degree i. Then, with probability at least $1 - O(n^{-13/27})$, G has diameter $O(\log n)$.*

Proof. Let $m = \sum_i i \cdot a_i n$ be the total number of copies. Let $N(k)$ denote the number of ways to pair up k copies. Thus, $N(k) = 0$ if k is odd, and $N(k) = (k-1)!! = \frac{k!}{(k/2)! \cdot 2^{k/2}}$ if k is even.

Let $P(U, s)$ be the probability that a given set U of $u = |U|$ copies has at most s edges leaving it. To bound $P(U, s)$ from above we observe that if at most s edges leave U, then at least $u - s$ copies of U must be matched amongst each other, and at least $m - u - s$ copies of $\overline{U}$ must be matched amongst each other (while the remaining $2s$ copies can be matched arbitrarily). Thus, there are at most

$\binom{u}{s}\binom{m-u}{s}N(u-s)N(m-u-s)N(2s)$ ways in which this could happen. Thus,

$$\begin{aligned} P(U,s) &\le \binom{u}{s}\binom{m-u}{s}\cdot N(2s)\cdot N(u-s)\cdot N(m-u-s)\cdot\frac{1}{N(m)} \\ &= \frac{u!\,(m-u)!\,(m/2)!\,(2s)!}{(s!)^3\,((u-s)/2)!\,((m-u-s)/2)!\,m!} \\ &=: Q(u,s). \end{aligned}$$

We claim that with probability at least $1-O(m^{-13/27})$, each set U of nodes with a total of $u \le m/2$ copies (i.e., $\sum_{v\in U}\text{degree}(v) \le u$) has at least $u/81$ edges leaving it. The theorem then follows by observing that the breadth-first tree vertex rooted at any vertex v will have at least $(1+1/81)^d$ vertices at depth d.

As we will be taking a Union Bound over all sets U, we need to bound how many distinct node sets U can contain exactly u copies. Because each node has degree at least 3, a set of at most u copies can contain at most $u/3$ nodes, and in fact, it can be seen that there are at most $\binom{m/3}{u/3}$ sets of exactly u copies. Then, we can bound

$$\binom{m/3}{u/3}Q(u,s) = \frac{(m/3)!\,u!\,(m-u)!\,(m/2)!\,(2s)!}{(u/3)!\,((m-u)/3)!\,(s!)^3\,((u-s)/2)!\,((m-u-s)/2)!m!}. \tag{10}$$

To simplify this expression further, we use the following version of Stirling's Approximation:

$$\sqrt{2\pi n}\cdot n^n e^{-n} e^{1/(12n+1)} \le n! \le \sqrt{2\pi n}\cdot n^n e^{-n} e^{1/12n}.$$

Substituting the upper bound in the numerator, and the lower bound in the denominator, we see that all the e^{-n} factors cancel out, while the product of all of the $e^{1/(12n+1)}$ and $e^{-1/12n}$ factors is bounded by e^2, which is less than the $(2\pi)^{-3/2}$ factor we obtain from the $\sqrt{2\pi}$ factors. The expression in Eq. (10) is thus bounded by

$$\sqrt{\frac{12m}{s^2(u-s)(m-u-s)}}\cdot\frac{(m/3)^{m/3}u^u\,(m-u)^{m-u}}{(u/3)^{u/3}\,((m-u)/3)^{(m-u)/3}\,s^{3s}\,((u-s)/2)^{(u-s)/2}}\cdot \frac{(m/2)^{m/2}(2s)^{2s}}{((m-u-s)/2)^{(m-u-s)/2}\,m^m}.$$

Grouping by powers of m, u, s, $u-s$, $m-u$, $m-u-s$, 2, and 3, this simplifies to

$$\sqrt{\frac{12m}{s^2(u-s)(m-u-s)}}\cdot\frac{u^{2u/3}\,(m-u)^{(2/3)(m-u)}\,2^s}{m^{m/6}\,s^s\,(u-s)^{(u-s)/2}\,(m-u-s)^{(m-u-s)/2}}. \tag{11}$$

Substituting $s = u/81$, and, using the fact $3 \le u \le m/2$ to bound the leading constant, the expression in Eq. (11) is bounded by

$$Ac^u\,\frac{u^{13u/81}\,(m-u)^{(2/3)(m-u)}}{m^{m/6}\,(m-82u/81)^{(m-82u/81)/2}} =: f(u)$$

where $A = 3^7\sqrt{2}/40$ and $c = (3^{164}/(2^{159}5^{40}))^{1/81}$.

We will now prove that $\ln f(u)$ is a concave function of u. Its first derivative is

$$\frac{d}{du}\ln f(u) = \frac{13}{81}\ln u - \frac{2}{3}\ln(m-u) + \frac{41}{81}\ln(m - 82u/81) + \ln c,$$

and its second derivative is

$$\begin{aligned}\frac{d^2}{du^2}\ln f(u) &= \frac{13}{81u} + \frac{2}{3(m-u)} - \frac{82\cdot 41}{81\cdot(81m-82u)}\\ &= \frac{13}{81u} + \frac{2}{81}\frac{506m-533u}{(81m-82u)(m-u)},\end{aligned}$$

which is positive for $1 \le u \le m/2$.

It follows that in any subinterval of $[1, m/2]$, both $\ln f(u)$ and $f(u)$ attain their maximum at one of the endpoints. In particular, we note that $f(3) = O(m^{-13/27})$, $f(12) = O(m^{-52/27})$, and

$$f(m/2) = O\left(\left(\tfrac{9}{2^{173/81}5^{40/81}}\right)^m\right) = o(0.93^m).$$

Thus, the probability that some set of $3 \le u \le m/2$ vertices has fewer than $u/81$ edges leaving it is at most

$$\begin{aligned}\sum_{u=3}^{m/2} f(u) &= \sum_{u=3}^{11} f(u) + \sum_{u=12}^{m/2} f(u)\\ &\le 9\cdot f(3) + m/2\cdot f(12)\\ &= O(m^{-13/27}) + O(m^{-25/27})\\ &= O(m^{-13/27}).\end{aligned}$$

Hence, with probability at least $1 - O(m^{-13/27})$, the diameter of the graph is $O(\log m) = O(\log n)$. □

5. *Generating Functions*

In this section, we use the formalism of generating functions [Wilf 1994] to express the results of Section 3 more succinctly, and complete the proof of Theorem 1.2. 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 3.7,

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

Using the generating function formalism, we can write

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

and from Eq. (7) we have

$$\begin{aligned} p_{\text{vis}}(t) &= \sum_k \frac{k a_k t^k}{t g'(t)} \left(\frac{g'(t)}{t g'(1)} \right)^k \\ &= \frac{1}{t g'(t)} \sum_k k a_k \left(\frac{g'(t)}{g'(1)} \right)^k \\ &= \frac{1}{t g'(1)} g' \left(\frac{g'(t)}{g'(1)} \right). \end{aligned} \tag{12}$$

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

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

which completes the proof of Theorem 1.2.

Our definition of "sober" degree sequences implies that the graph is with high probability connected, so that every copy is eventually added to the queue. For other degree sequences, Molloy and Reed Molloy and Reed [1995, 1998] established that with high probability 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 $g^{\text{obs}}(z)$ is then given by an integral from t_0 to 1, where t_0 is the time at which the giant component has with high probability been completely exposed; this is the time at which $c_{\text{unto}}(t) = c_{\text{unex}}(t)$, namely the largest root less than 1 of the equation

$$\sum_j j a_j t^j = t^2 \sum_j j a_j. \tag{13}$$

6. *Examples*

Using the machinery developed above, we now derive the observed degree sequences for two simple and well-known random graph degree sequences. Although both examples produce a power-law degree sequence with exponent $\alpha = 1$, other underlying sequences can produce different exponents.

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 Eq. (1), we derive the generating function for the observed degree sequence:

$$g^{\text{obs}}(z) = z\delta \cdot \int_0^1 t^{\delta-1}(1-(1-z)t^{\delta(\delta-2)})^{\delta-1}\mathrm{d}t. \tag{14}$$

This integral can be expressed in terms of the hypergeometric function ${}_2F_1$ [Seaborn 1991]. 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 Eq. (14), $a = \delta - 1$, $b = \delta(\delta-2)$, and $c = 1-\delta$ (note $a > -1$ and $b > 0$ since $\delta > 2$) giving

$$g^{\text{obs}}(z) = z \cdot {}_2F_1\left(\frac{1}{\delta-2}, 1-\delta; 1+\frac{1}{\delta-2}; 1-z\right). \tag{15}$$

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, Eq. (15) becomes

$$\begin{aligned} g^{\text{obs}}(z) &= z \cdot \frac{\Gamma(1+\frac{1}{\delta-2})\,\Gamma(\delta)}{\Gamma(\delta+\frac{1}{\delta-2})} \cdot {}_2F_1\left(\frac{1}{\delta-2}, 1-\delta; 1-\delta; z\right) \\ &= z \cdot \frac{\Gamma(\delta)}{\Gamma(\delta+\frac{1}{\delta-2})\,(\delta-2)} \sum_{m=0}^{\infty}\Gamma\left(m+\frac{1}{\delta-2}\right)\frac{z^m}{m!}. \end{aligned}$$

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:

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

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 [2005] 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 a result about the observed degree distribution of random graphs with Poisson degree distribution. (Recall that the degree distribution of $G(n, p = \delta/n)$ is with overwhelmingly high probability within $o(n)$ of a Poisson distribution with mean δ.) The generating function for a degree distribution that is Poisson with mean δ is $g(z) = \mathrm{e}^{-\delta(1-z)}$, and the generating function for the observed degree sequence is

$$\begin{aligned} g^{\text{obs}}(z) &= z\delta \cdot \int_{t_0}^{1} \mathrm{e}^{-\delta(1-t)} \mathrm{e}^{-\delta(1-z)\mathrm{e}^{-\delta(1-\mathrm{e}^{-\delta(1-t)})}} \mathrm{d}t \\ &= z \int_{0}^{1-t_0} \mathrm{e}^{-\delta(1-z)\mathrm{e}^{-\delta y}} \mathrm{d}y. \end{aligned} \tag{16}$$

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

This integral can be expressed in terms of the exponential integral function $\mathrm{Ei}(z)$ [Spanier and Oldham 1987] and the incomplete Gamma function $\Gamma(a, z)$, which are defined as

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

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

Then, with the integral

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

and the Taylor series

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

taking $a = -\delta(1-z)$ and $b = -\delta$ as in Eq. (16) gives

$$\begin{aligned} g^{\text{obs}}(z) &\approx \tfrac{z}{\delta} \left(\mathrm{Ei}\Big[-\delta(1-z) \Big] - \mathrm{Ei}\left[-\delta \mathrm{e}^{-\delta(1-t_0)}(1-z) \right] \right) \\ &= \textstyle\sum_{m=0}^{\infty} \frac{z^{m+1}}{\delta m!} \left(\Gamma(m, \delta \mathrm{e}^{-\delta(1-t_0)}) - \Gamma(m, \delta) \right). \end{aligned} \tag{17}$$

Thus, the coefficients of the observed degree sequence are

$$a_{m+1}^{\text{obs}} = \tfrac{1}{\delta m!} \int_{\delta \mathrm{e}^{-\delta(1-t_0)}}^{\delta} \mathrm{e}^{-x} x^{m-1} \, \mathrm{d}x. \tag{18}$$

Now, t_0 approaches $e^{-\delta}$ in the limit of large δ, and for $m \lesssim \delta$, the integral of Eq. (18) coincides almost exactly with the full Gamma function $\Gamma(m)$ since it contains the peak of the integrand. Specifically, Clauset and Moore [2005] 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. [2003].

7. *Conclusions*

We have established rigorously that single-source traceroute sampling is biased, thus formally verifying the empirical observations of Lakhina et al. [2003], and we have calculated the precise nature of that bias for a broad class of random graphs. Recently, Cohen et al. [2007] and, independently, M. Bousquet-Mélou (personal communication) used our machinery to show asymptotically that when the graph has a power-law degree distribution, at the highest degrees the observed exponent converges to the underlying one. However, vertices of low and moderate degree have a large effect on numerical estimates of the exponent, especially in finite-sized graphs or when the average degree of the underlying graph is relatively large. For example, Petermann and De Los Rios [2004] and Clauset and Moore [2005] both demonstrated experimentally that traceroute sampling tends to significantly underestimate the true exponent α for precisely this reason.

All of the analytical results to date assume essentially uniformly random graph models for the Internet. This is clearly a simplifying assumption. Indeed, there is no question as to whether such graphs are good models of the Internet: they are clearly not. As observed for instance by Alderson et al. [2005], accurate Internet models should take its required function and economic constraints into account, and will thus inherently lead to more clustering than uniformly random models. However, such models are beyond the purview of current analytical techniques, and a rigorous exploration of the inherent bias of sampling techniques like traceroute is interesting in its own right.

Several recent articles, published since the appearance of the conference version of this article, continue an analysis of the impact of traceroute sampling on the exploration of random graphs. Blondel et al. [2007] use mean-field approximations to heuristically calculate the distance distribution of nodes in random networks, as well as the fraction of edges lying on at least one shortest path from s to some node. They observe and analyze in depth an oscillating behavior for the latter quantity. Dall'Asta et al. [2006] use mean-field approximations and simulations to explore the probabilities of vertex and edge detection, and relate these probabilities to the betweenness of vertices and edges.

Despite several simulation-based studies, there is so far no rigorous mathematical analysis of the benefit of using traceroutes from multiple sources. Several empirical studies suggest that a small number of additional sources in IP-level mapping studies

have at best a small marginal utility [Barford et al. 2001; Pansiot and Grad 1998]. However, numerical studies of power-law random graphs by Clauset and Moore [2005] and Guillaume et al. [2006], and an analytical study of Poisson random graphs by Dall'Asta [2007], show that additional sources can have significant utility in sufficient numbers. For 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. A rigorous analysis of this result is complicated by the fact that the events corresponding to a fixed edge appearing in different BFS trees are highly correlated. We leave the generalization of our results to traceroute sampling with multiple sources as future work.

While our work rigorously establishes the bias introduced by single-source BFS tree sampling of random graphs, it does not prescribe ways to mitigate or invert the bias. Several possible approaches suggest themselves. Combining IP-level route data from a large number of geographically distributed sources will likely improve the accuracy of our Internet maps. However, many practical issues related to the ownership and management of IP-level routers will, for the foreseeable future, strongly limit the degree of coverage achievable in this way, even for reasonably large-scale sampling projects like NetDimes [Shavitt and Shir 2005]. Because the current level of coverage may not be sufficient to overcome the inherent bias of traceroute sampling, another possibility is to employ sophisticated inference or machine learning techniques that rely mainly on data currently accessible to researchers. For instance, Flaxman and Vera [2007] recently proposed a new estimator of node degrees using insights from multiple-capture census techniques in biology. Under certain strong assumptions, this technique provably reduces the sampling bias.

A very desirable result would be a way to invert Theorem 1.2, and derive $g(z)$ from $g^{\text{obs}}(z)$. This would allow us to explicitly 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 $g^{\text{obs}}(z)$ is invertible, and the complexity of our expression for $g^{\text{obs}}(z)$ makes such an inversion appear quite difficult. However, it remains an exciting and challenging direction for future work.

ACKNOWLEDGMENT. We thank André Allavena, Luca Dall'Asta, Jon Kleinberg, and Tracy Conrad for helpful conversations. C.M. also thanks Rosemary Moore for providing a larger perspective.

REFERENCES

ALDERSON, D., LI, L., WILLINGER, W., AND DOYLE, J. 2005. Understanding Internet topology: Principles, models, and validation. *IEEE/ACM Trans. Netw. 13*, 1205–1218.

AMINI, L., SHAIKH, A., AND SCHULZRINNE, H. 2002. Issues with inferring Internet topological attributes. In *SPIE IT-Com*.

BARFORD, P., BESTAVROS, A., BYERS, J., AND CROVELLA, M. 2001. On the marginal utility of network topology measurements. In *SIGCOMM Internet Measurement Workshop*.

BLONDEL, V., GUILLAUME, J.-L., HENDRICKX, J., AND JUNGERS, R. 2007. Distance distribution in random graphs and application to network exploration. *Phys. Rev. E 76*, 066101.

BOLLOBÁS, B. 2001. *Random Graphs*, 2nd ed. Cambridge University Press, Cambridge.

BOLLOBÁS, B., AND CHUNG, F. 1988. The diameter of a cycle plus a random matching. *SIAM J. Disc. Math. 1*, 328–333.

CHEN, Q., CHANG, H., GOVINDAN, R., JAMIN, S., SHENKER, S., AND WILLINGER, W. 2002. The origin of power laws in Internet topologies revisited. In *Proceedings of the 21st ACM SIGCOMM Conference*. ACM, New York.

CLAUSET, A., AND MOORE, C. 2005. Accuracy and scaling phenomena in Internet mapping. *Phys. Rev. Lett. 94*, 018701.

COHEN, R., GONEN, M., AND WOOL, A. 2007. Bounding the bias of tree-like sampling in ip topologies. In *Proceedings of the European Conference on Complex Systems (ECCS)*.

DALL'ASTA, L. 2007. Dynamic exploration of networks: From general principles to the traceroute process. *Europ. Phys. J. B 60*, 123–133.

DALL'ASTA, L., ALVAREZ-HAMELIN, I., BARRAT, A., VÁZQUEZ, A., AND VESPIGNANI, A. 2006. Exploring networks with traceroute-like probes: Theory and simulations. *Theoret. Comput. Sci. 355*, 6–24.

ERDŐS, P., AND RÉNYI, A. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci. 5*, 17–61.

FABRIKANT, A., KOUTSOUPIAS, E., AND PAPADIMITRIOU, C. 2002. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proceedings of the 29th International Colloguium on Automata, Languages and Programming*. 110–122.

FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. 1999. On power-law relationships of the Internet topology. In *Proceedings of the 18th ACM SIGCOMM Conference*. ACM, New York. 251–262.

FLAXMAN, A., AND VERA, J. 2007. Bias reduction in traceroute sampling: Towards a more accurate map of the internet. In *Proceedings of the 5th Workshop on Algorithms and Models for the Web-Graph (WAW2007)*.

GOVINDAN, R., AND TANGMUNARUNKIT, H. 2000. Heuristics for Internet map discovery. In *Proceedings of the 19th ACM SIGCOMM Conference*. ACM, New York.

GUILLAUME, J.-L., LATAPY, M., AND MAGONI, D. 2006. Relevance of massively distributed explorations of the internet topology: Qualitative results. *Comput. Netw. 50*, 3197–3224.

HOEFFDING, W. 1963. Probability inequalities for sums of bounded random variables. *J. ASA 58*, 13–30.

KIM, J. H. 2006. Poisson cloning model for random graphs. Preprint.

KLEINBERG, J., KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. The web as a graph: Measurements, models and methods. In *Proceedings of the International Conference on Combinatorics and Computing*.

LAKHINA, A., BYERS, J., CROVELLA, M., AND XIE, P. 2003. Sampling biases in IP topology measurements. In *Proceedings of the 22nd IEEE INFOCOM Conference*. IEEE Computer Society Press, Los Alamitos.

LEGUAY, J., LATAPY, M., FRIEDMAN, T., AND SALAMATIAN, K. 2007. Describing and simulating Internet routes. *Comput. Netw. 51*, 2067–2087.

MCDIARMID, C. 1998. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, Eds. Springer-Verlag, Berlin, Germany, 195–247.

MIHAIL, M., PAPADIMITRIOU, C., AND SABERI, A. 2003. On certain connectivity properties of the Internet topology. In *Proceedings of the 35th ACM Symposium on Theory of Computing*. ACM, New York.

MOLLOY, M., AND REED, B. 1995. A critical point for random graphs with a given degree sequence. *Rand. Struct. Algor. 6*, 161–180.

MOLLOY, M., AND REED, B. 1998. The size of the largest component of a random graph on a fixed degree sequence. *Combin. Probab. Comput. 7*, 295–306.

MOTWANI, R., AND RAGHAVAN, P. 1990. *Randomized Algorithms*. Cambridge University Press. Cambridge.

PANSIOT, J.-J., AND GRAD, D. 1998. On routers and multicast trees in the Internet. *ACM SIGCOMM Commun. Rev. 28*, 41–50.

PETERMANN, T., AND DE LOS RIOS, P. 2004. Exploration of scale-free networks: Do we measure the real exponents? *Euro. Phys. J. B 38*, 201–204.

SEABORN, J. 1991. *Hypergeometric Functions and Their Applications*. Springer-Verlag, Berlin, Germany.

SHAVITT, Y., AND SHIR, E. 2005. Dimes: Let the Internet measure itself. *ACM SIGCOMM Comput. Commun. Rev. 35*, 71–74.

SPANIER, J., AND OLDHAM, K. 1987. The exponential integral ei(x) and related functions. In *An Atlas of Functions*. Hemisphere, Chapter 37, 351–360.

SPRING, N., MAHAJAN, R., WETHERALL, D., AND ANDERSON, T. 2004. Measuring ISP topologies with rocketfuel. *IEEE/ACM Trans. Netw. 12*, 2–16.

TANGMUNARUNKIT, H., GOVINDAN, R., SHENKER, S., AND ESTRIN, D. 2001. The impact of policy on Internet paths. In *Proceedings of the 20th IEEE INFOCOM Conference*. IEEE Computer Society Press,

Los Alamitos.
VIGER, F., AUGUSTIN, B., CUVELLIER, X., MAGNIEN, C., LATAPY, M., FRIEDMAN, T., AND TEIXEIRA, R. 2006. Detection, understanding, and prevention of traceroute measurement artifacts. In *Proceedings of the 6th Internet Measurement Conference (IMC'06)*.
WILF, H. 1994. *Generatingfunctionology*. Academic Press, Orlando, FL.
WORMALD, N. 1999. Models of random regular graphs. In *London Math. Soc. Lecture Note Series*, J. Lamb and D. Preece, Eds. Vol. 276. Cambridge University Press, Cambridge. 239–298.

RECEIVED JULY 2007; REVISED JANUARY 2008; ACCEPTED FEBRUARY 2008