Internet Mathematics Vol. 3, No. 2: 147-152

Random Walks with Lookahead on Power Law Random Graphs

Milena Mihail, Amin Saberi, and Prasad Tetali

Abstract. We show that in power law random graphs, almost surely, the expected rate at which a random walk with lookahead discovers the nodes of the graph is sublinear.

Searching a graph by simulating a random walk is a natural way to abstract web crawling [Cooper and Frieze 02, Cooper and Frieze 03a]. Recently, the random walk simulation method has also been proposed to search P2P networks [Lv et al. 02, Chawathe et al. 03, Gkantsidis et al. 04]. Therefore, it is important to characterize the rate at which a random walk discovers the vertices of large, sparse graphs. Strong bounds indicating behavior similar to coupon collection have been obtained by Cooper and Frieze who show that, almost surely, the expected cover time of a random d-regular graph is $\frac{d-1}{d-2} n \log n$ [Cooper and Frieze 03b, Cooper and Frieze 05], and that, almost surely, the expected cover time of a random scale free graph, in the model of growth with preferential attachment, is $\frac{2d}{d-1} n \log n$, where d is the average degree [Cooper and Frieze 04]. Since the degrees of the World Wide Web are known to follow heavy-tailed statistics, it is important to study random graph models resulting in heavy-tailed degree distributions.

In this paper we formalize a common practice of crawling, namely *lookahead*. In a lookahead-1 scenario, when a crawler visits a node v, he is assumed to also discover all the neighbors of v. This is particularly efficient to implement in a sparse network by having each node keep a copy of the indices of all his neighbors. The resulting replication overhead is proportional to the number of

1542-7951/06 $0.50 per page

edges in the network, which for sparse networks is linear. A further practice is lookahead 2, where, for every visited node v, the random walk is assumed to also discover all the neighbors of v and all the neighbors' neighbors, $N_2(v)$ (see also [Manku et al. 04] for an application of lookahead 2 in routing).

We show that, in the power law random graph model [Aiello et al. 00], a.s. (for all but a vanishingly small fraction of the graphs), the expected time at which a random walk with lookahead discovers the graph is sublinear, in particular, much faster than even coupon collection. Intuitively, the reason for these savings is that the stationary distribution of the random walk biases the search towards high-degree nodes that yield a large amount of information about their neighbors. Therefore, in some sense, our results suggest that the practice of lookahead explores the heavy-tailed statistics of the network to sharply improve the performance of the search algorithm.

The power law random graph model is as follows. Given n and ϵ, $0<\epsilon<1$, we first generate degrees d_i, $1\le i\le n$, independently, according to the distribution $\Pr[d_i=x] \simeq \frac{c}{x^{2+\epsilon}}$, $d_{\min}\le x\le\sqrt{n}$, where c is a normalizing constant. We then consider $D=\sum_{i=1}^{n} d_i$ minivertices that correspond to vertices in the natural way. Finally, we consider a random perfect matching over D and, for every edge in the matching between a minivertex corresponding to vertex i and a minivertex corresponding to vertex j, we add a distinct edge connecting vertex i with vertex j. This is a multigraph with self loops; we maintain multiple edges and self loops for analytic convenience. Gkantsidis et al. [Gkantsidis et al. 03] show that, for a large enough constant $d_{\min}$, this random graph has conductance $\Omega(1)$, almost surely. Following the standard theory of mixing times, this implies that, after $O(\log n)$ steps, the distribution of the random walk is within variation distance $O(\text{poly}^{-1}(n))$ from its stationary distribution. Now, standard coupon collection arguments suggest an expected cover time of $O(n\log^2 n)$. For a description of the coupon collector's problem and Chernoff bounds, see [Alon and Spencer 00, Motwani and Raghavan 95]. Our results are Theorems 1.1 and 1.2 below.

Theorem 1.1. *For any δ, $0<\delta<\frac{1}{2}$, the expected number of simulation steps for a random walk (starting from an arbitrary distribution) with lookahead 1 to discover $\Omega(n^{1-\epsilon(\frac{1}{2}-\delta)})$ vertices is $O(n^{\frac{1}{2}+\delta}\log n)$, a.s.*

Theorem 1.2. *For any δ, $0<\delta<\frac{1}{2}$, the expected number of simulation steps for a random walk (starting from an arbitrary distribution) with lookahead 2 to discover $\Omega(n^{1-2\epsilon(\frac{1}{2}-\delta)-\delta})$ vertices is $O(n^{\epsilon(\frac{1}{2}-\delta)}\log n)$, a.s.*

The proofs of Theorems 1.1 and 1.2 follow from the rapid mixing of the random walk and the structural Lemmas 1.6, 1.7, and 1.8, given below. We also need

Facts 1.3, 1.4, and 1.5. The form of the Chernoff bounds quoted below is from page 29 of [Spencer 93].

Fact 1.3. *$D=\sum_{i=1}^{n} d_i = O(n)$, a.s.*

Proof. The mean of D is dn, for some constant d, and the variance of D can be computed to be $\sigma=\Theta(n^{\frac{3}{4}-\frac{\epsilon}{4}})$. Now, using the tail inequality in Theorem A.1.19 of [Alon and Spencer 00], we get that, for every $\alpha \leq O(\sigma/\sqrt{n})$, $\Pr[D - dn > \alpha\sigma] < e^{-\frac{\alpha^2}{4}}$. If we pick $\alpha = n^{\frac{1}{4}-\frac{\epsilon}{4}}$, we get the desired bound. □

Fact 1.4. *There are $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$ vertices of degree $n^{\frac{1}{2}-\delta}$, a.s.*

Proof. We first compute that

$$\Pr[d_i \geq n^{\frac{1}{2}-\delta}] = \sum_{x=n^{\frac{1}{2}-\delta}}^{\sqrt{n}} \frac{c}{x^{2+\epsilon}} = \Omega(n^{-(\frac{1}{2}-\delta)(1+\epsilon)}).$$

Hence, the expected number of large vertices is $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$, and, by the Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$ large vertices, a.s. □

Henceforth we will call the vertices guaranteed by Fact 1.4 *large* vertices.

Fact 1.5. *There are $\Omega(n)$ vertices of degree $d_{\min}$, a.s.*

Proof. The probability that a vertex has degree $d_{\min}$ is a constant, therefore the expected number of vertices of degree $d_{\min}$ is $\Omega(n)$, and by the Chernoff bounds, there are $\Omega(n)$ vertices of degree $d_{\min}$, a.s. □

Lemma 1.6. *Each large vertex has $\Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$ edges incident to distinct large vertices, a.s.*

Proof. First, for $k=\Theta(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)})$, we bound the probability that a large vertex has at least $k+1$ edges incident to other distinct large vertices, conditioned on the fact that it has at least k edges incident to distinct large vertices. This can be bounded by $\frac{\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})\Omega(n^{\frac{1}{2}-\delta})-kn^{\frac{1}{2}}}{\Theta(n)} = \Omega(n^{-\epsilon(\frac{1}{2}-\delta)})$. Now, we can see that, over k edges incident to the large vertex, the expected number of edges incident to distinct large vertices is $\Omega(kn^{-\epsilon(\frac{1}{2}-\delta)}) = \Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$, and, by the Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$ edges incident to distinct large vertices, a.s. □

Lemma 1.7. *Each large vertex has $\Omega(n^{\frac{1}{2}-\delta})$ edges incident to vertices of degree $d_{\min}$, a.s.*

Proof. Let d be the degree of a large vertex. First notice that, if we condition on the fact that the first $d-1$ edges incident to the large vertex have their other endpoint incident to vertices of degree $d_{\min}$, the probability that the dth edge has its other endpoint incident to a vertex of degree $d_{\min}$ is $\Omega(1)$. Now, it can be seen that the expected number of edges incident to the large vertex that have their other endpoint incident to a vertex of degree $d_{\min}$ is $\Omega(n^{\frac{1}{2}-\delta})$, and, by the Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-\delta})$ such edges, a.s. □

Lemma 1.8. *For every large vertex v, $|N_2(v)| = \Omega(n^{1-2\epsilon(\frac{1}{2}-\delta)-\delta})$, a.s.*

Proof. By Lemma 1.6, v has $\Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$ distinct large neighbors. By Lemma 1.7, each large neighbor has $\Omega(n^{\frac{1}{2}-\delta})$ edges incident to vertices of degree $d_{\min}$, for a total of $\Omega(n^{1-2\epsilon(\frac{1}{2}-\delta)-\delta})$ edges incident to vertices of degree $d_{\min}$. But, each vertex of degree $d_{\min}$ can take at most $d_{\min}$ edges, hence there are $\Omega(n^{1-2\epsilon(\frac{1}{2}-\delta)-\delta})$ distinct vertices of degree $d_{\min}$ in $N_2(v)$. □

Proof of Theorem 1.1. By the rapid mixing shown in [Gkantsidis et al. 03], $O(\log n)$ simulation steps get a sample from a distribution arbitrarily close to the stationary. By Facts 1.3 and 1.4, we can compute the stationary probability of the set of large vertices as $\Omega(n^{-\epsilon(\frac{1}{2}-\delta)})$, and hence, in expected time $O(n^{\epsilon(\frac{1}{2}-\delta)})$, we get a large vertex. Now, by coupon collection we will get $\Omega(n^{\frac{1}{2}+\delta-\epsilon(\frac{1}{2}-\delta)})$ distinct large vertices in expected time $O(n^{\frac{1}{2}+\delta}\log n)$. Let L be the set of sampled large vertices. By Lemma 1.7, L has $\Omega(n^{1-\epsilon(\frac{1}{2}-\delta)})$ edges incident to vertices with degree $d_{\min}$, and since each vertex with degree $d_{\min}$ can be incident to at most $d_{\min}$ distinct large vertices, we get $\Omega(n^{1-\epsilon(\frac{1}{2}-\delta)})$ distinct vertices of degree $d_{\min}$. □

Proof of Theorem 1.2. The expected time to see one large vertex is $O(n^{\epsilon(\frac{1}{2}-\delta)})$. Now, Theorem 1.2 follows from the size of the 2-neighborhood of this large vertex established in Lemma 1.8. □

Finally, we should mention that the first reference to the potential power of lookahead in searching power law graphs is due to [Adamic et al. 01]. However, their analytic results refer to a graph with all its crucial random variables behaving as their expected values. In particular, the theorem in [Adamic et al. 01], namely cover time $O(\log^2 n)$ for a random walk with lookahead 2, does not seem to hold here. This is because, a.s., the graph will have $\Omega(n)$ small degree vertices with their entire 2-neighborhoods also consisting of small degree vertices; hence, we need $\Omega(n)$ sample points to discover this set.

References

[Adamic et al. 01] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. "Search in Power Law Networks." *Phy. Rev. E* 64 (2001), 46135.

[Aiello et al. 00] W. Aiello, F. Chung, and L. Lu. "A Random Graph Model for Power Law Graphs." In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pp. 171–180. Los Alamitos, CA: IEEE Press, 2000.

[Alon and Spencer 00] N. Alon and J. Spencer. *The Probabilistic Method*, Second edition. Hoboken, NJ: John Wiley & Sons, 2000.

[Chawathe et al. 03] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. "Making Gnutella-like P2P Systems Scalable." In *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 407–418. New York: ACM Press, 2003.

[Cooper and Frieze 02] C. Cooper and A. Frieze. "Crawling on Web Graphs." In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 419–427. New York: ACM Press, 2002.

[Cooper and Frieze 03a] C. Cooper and A. Frieze. "Crawling on Simple Models of Web Graphs." *Internet Mathematics* 1:1 (2003), 57–90.

[Cooper and Frieze 03b] C. Cooper and A. Frieze. "The Cover Time of Sparse Random Graphs." In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 148–157. Philadelphia, PA: SIAM, 2003.

[Cooper and Frieze 04] C. Cooper and A. Frieze. "The Cover Time of the Preferential Attachment Graph." Preprint, 2004.

[Cooper and Frieze 05] C. Cooper and A. Frieze. "The Cover Time of Random Regular Graphs." *SIAM Journal of Discrete Mathematics* 18 (2005), 728–740.

[Gkantsidis et al. 03] C. Gkantsidis, M. Mihail, and A. Saberi. "Conductance and Congestion in Power Law Graphs." In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 148–159. New York: ACM Press, 2003.

[Gkantsidis et al. 04] C. Gkantsidis, M. Mihail, and A. Saberi. "Random Walks in Peer-to-Peer Networks." In *INFOCOM 2004: Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, pp. 120–131. Los Alamitos, CA: IEEE Press, 2004.

[Lv et al. 02] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. "Search and Replication in Unstructured Peer-to-Peer Networks." In *Proceedings of the 16th International Conference on Supercomputing*, pp. 84–95. New York: ACM Press, 2002.

[Manku et al. 04] G. S. Manku, M. Naor, and U. Wieder. "Know Thy Neighbor's Neighbor: The Power of Lookahead in Randomized P2P Networks." In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 54–63. New York: ACM Press, 2004.

[Motwani and Raghavan 95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 1995.

[Spencer 93] J. Spencer. *Ten Lectures on the Probabilistic Method*, Second edition. Philadelphia: SIAM, 1993.

Milena Mihail, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 (mihail@cc.gatech.edu)

Amin Saberi, Department of Management Science and Engineering and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 (saberi@stanford.edu)

Prasad Tetali, School of Mathematics and College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 (tetali@math.gatech.edu)

Received July 29, 2005; accepted November 7, 2005.