Random Walks with Lookahead in Power Law Random Graphs

Milena Mihail   Amin Saberi   Prasad Tetali
Georgia Institute of Technology
Email: {mihail, saberi}@cc.gatech.edu tetali@math.cc.gatech.edu

Abstract—**We show that in power law random graphs, a.s., 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 [5]. Recently, the random walk simulation method has been also proposed to search P2P networks [11], [4], [10]. 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 [6] [7] who show that, a.s., the expected cover time of a random d-regular graph is $\frac{d-1}{d-2}n\log n$, and by [8] who show that, a.s., 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. Since the degrees of the WWW 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 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 [12] for an application of lookahead 2 in routing).

We show that, in the power law random graph model [2], 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 (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 which 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 which 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. [9] show that, for a large enough constant $d_{\min}$, this random graph has conductance $\Omega(1)$, almost surely. Following standard theory of mixing times, this implies that, after $O(\log n)$ steps, the distribution of the random walk is within variation distance $O(\mathrm{poly}^{-1}(n))$ from its stationary distribution. Now standard coupon collection arguments suggest expected cover time $O(n\log^2 n)$. Our results are Theorems 1 and 2 below.

Theorem 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 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 and 2 follow from the rapid mixing of the random walk and the structural Lemmas 6, 7 and 8 below. We also need Facts 3, 4 and 5. The form of Chernoff bounds quoted is from page 29 of [13].

Fact 3: $D=\sum_{i=1}^{n}d_i=O(n)$, a.s.

PROOF OF FACT 3. 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 , page 270 of [3], we get that, for every $\alpha\le$

$O(\sigma/\sqrt{n})$, $\Pr[D - dn > \alpha\sigma] < e^{-\frac{\alpha^2}{4}}$. If we pick $\alpha = n^{\frac{1}{4}-\frac{\epsilon}{4}}$ we get the desired bound.

Fact 4: There are $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$ vertices of degree $n^{\frac{1}{2}-\delta}$, a.s. We will henceforth call these vertices large.

PROOF OF FACT 4. We first compute that $\Pr[d_i \geq n^{\frac{1}{2}-\delta}] = \sum_{x=d_{n^{\frac{1}{2}-\delta}}}^{\sqrt{n}} \frac{c}{x^{2+\epsilon}} = \Omega(n^{-(\frac{1}{2}-\delta)(1+\epsilon)})$. Hence the expected number of large vertices is $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$ and, by Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)+\delta})$ large vertices, a.s.

Fact 5: There are $\Omega(n)$ vertices of degree $d_{\min}$, a.s.

PROOF OF FACT 5. 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 Chernoff bounds, there are $\Omega(n)$ vertices of degree $d_{\min}$, a.s.

Lemma 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 OF LEMMA 6. We first bound the probability that the large vertex has at least $k+1$ edges incident to distinct large vertices, conditioned on the fact that it has at least k edges incident to distinct large vertices, for $k = \Theta(n^{\frac{1}{2}-\epsilon(\frac{1}{2}-\delta)})$. 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 Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$ edges incident to distinct large vertices, a.s.

Lemma 7: Each large vertex has $\Omega(n^{\frac{1}{2}-\delta})$ edges incident to vertices of degree $d_{\min}$, a.s.

PROOF OF LEMMA 7. 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 d-th 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 Chernoff bounds, there are $\Omega(n^{\frac{1}{2}-\delta})$ such edges, a.s

Lemma 8: For every large vertex v, $N_2(v) = \Omega(n^{1-2\epsilon(\frac{1}{2}-\delta)-\delta})$, a.s.

PROOF OF LEMMA 8. By Lemma 6, v has $\Omega(n^{\frac{1}{2}-2\epsilon(\frac{1}{2}-\delta)})$ distinct large neighbors. By Lemma 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. By the rapid mixing shown in [9], $O(\log n)$ simulation steps get a sample from a distribution arbitrarily close to the stationary. By Lemma 4 and Fact 3, 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 by Lemma 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 2. The expected time to see one large vertex is $O(n^{\epsilon(\frac{1}{2}-\delta)})$. Now Theorem 2 follows from the size of the size of the 2-neighborhood of this large vertex established in Lemma 8.

Finally, we should mention that the first reference to the potential power of lookahead in searching power law graphs is due to [1]. However, their analytic results refer to a graph with all its crucial random variables behaving as their expected values. In particular, the main theorem in [1], namely cover time $O(\log^2 n)$ for random walk with lookahead 2, almost surely does not hold in a power law random graph. 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

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

[2] Aiello, W., Chung, F.R.K. and Lu, L., "A Random Graph Model for Power Law Graphs", FOCS 2000, pp. 171-180.

[3] Alon, N. and Spencer, J., The Probabilistic Method, John Wiley & Sons, 2nd Edition, (2000).

[4] Chawathe, Y., Ratnasamy, S., Breslau, L. Lanham, N. and Shenker, S., "Making Gnutella-like Networks Scalable", Sigcomm 2003, pp 407-418.

[5] Cooper, C. and Frieze, A., "Crawling on Web Graphs", STOC 2002, pp 419-427, Internet Mathematics, Vol 1, pp 57-90.

[6] Cooper, C. and Frieze, A., "The Cover Time of Sparse Random Graphs", SODA 2003, pp 148-157.

[7] Cooper, C. and Frieze, A., "The Cover Time of Random Regular Graphs ", preprint, (2004).

[8] Cooper, C. and Frieze, A., "The Cover Time of the Preferential Attachment Graph", preprint, (2004).

[9] Gkantsidis, C., Mihail, M. and Saberi, A., "Conductance and Congestion in Power Law Graphs", Sigmetrics 2003.

[10] Gkantsidis, C., Mihail, M. and Saberi, A., "On the Random Walk Method for PP2 Networks", Proceedings of Infocom 2004, pp 148-159.

[11] Lv, Q., Cao, P., Cohen, E., Li, K. and Shenker, S., "Search and Replication in Unstructured P2P Networks", Supercomputing 2002, pp 84-95.

[12] Manku, G.S., Naor, M. and Wieder, U., "Know Thy Neighbor's Neighbor: The Power of Lookahead in Randomized P2P Networks", STOC 2004.

[13] Spencer, J., Ten Lectures on the Probabilistic Method, SIAM, (1987).