A fast algorithm to find all high degree vertices in graphs with a power law degree sequence

Colin Cooper, Tomasz Radzik, and Yiannis Siantos

Department of Informatics, King's College London, UK

Abstract. We develop a fast method for finding all high degree vertices of a connected graph with a power law degree sequence. The method uses a biassed random walk, where the bias is a function of the power law c of the degree sequence.

Let $G(t)$ be a t-vertex graph, with degree sequence power law $c \geq 3$ generated by a generalized preferential attachment process which adds m edges at each step. Let S_a be the set of all vertices of degree at least t^a in $G(t)$. We analyze a biassed random walk which makes transitions along undirected edges $\{x, y\}$ proportional to $(d(x)d(y))^b$, where $d(x)$ is the degree of vertex x and $b > 0$ is a constant parameter. Choosing the parameter $b = (c-1)(c-2)/(2c-3)$, the random walk discovers the set S_a completely in $\widetilde{O}(t^{1-2ab(1-\epsilon)})$ steps with high probability. The error parameter ϵ depends on c, a and m. We use the notation $\tilde{O}(x)$ to mean $O(x \log^k x)$ for some constant $k > 0$.

The cover time of the entire graph $G(t)$ by the biassed walk is $\widetilde{O}(t)$. Thus the expected time to discover all vertices by the biassed walk is not much higher than in the case of a simple random walk $\Theta(t \log t)$.

The standard preferential attachment process generates graphs with power law $c = 3$. Choosing search parameter $b = 2/3$ is appropriate for such graphs. We conduct experimental tests on a preferential attachment graph, and on a sample of the underlying graph of the WWW with power law $c \sim 3$ which support the claimed property.

1 Introduction

Many large networks have a heavy tailed degree sequence. Thus, although the majority of the vertices have constant degree, a very distinct minority have very large degrees. This particular property is the significant defining feature of such graphs. A log-log plot of the degree sequence breaks naturally into three parts. The lower range (small constant degree) where there may be curvature, as the power law approximation is incorrect. The middle range, of large but well represented vertex degrees, which give the characteristic straight line log-log plot of the power law coefficient. In the upper tail, where the sequence is far from concentrated, the plot is a spiky mess. See for example Figure 1 (the degree

sequence of a simulated preferential attachment graph with $m = 4$ edges added at each step) and Figure 3 (the degree sequence of the underlying graph of a sample of the WWW). In both cases the x-axis is $a = \log d / \log t$, where d is vertex degree, and t is the size of the graph.

In our work we focus on sampling the higher degree vertices, in both the middle range and upper tail. Our aim is to find *all these vertices*, and we propose a provably efficient method of obtaining those vertices in sub-linear time using a weighted random walk. One reason for finding all the higher degree vertices is that the upper tail is not concentrated, so no sub-sample will be representative. We consider a weighted random walk because, as there are few vertices even in the middle range, a simple random walk may take too long to obtain a statistically significant sample. Coupled with this is the impression that in many networks, for example the WWW, it is the high degree vertices which are important, both as hubs and authorities, and for pagerank calculations.

Previous work on efficient sampling of network characteristics arises in many areas. In the context of search engine design, studies in optimally sampling the URL crawl frontier to rapidly sample (e.g.) high pagerank vertices, based on knowledge of vertex degree in the current sample, can be found in e.g. [3].

Within the random graph community, *traceroute sampling* was used to estimate cumulate degree distributions; and methods of removing the high degree bias from this process were studied in e.g. [1], [10]. Another approach, analysed in [6], is the *jump and crawl* method to find (e.g.) all very high degree vertices. The method uses a mixture of uniform sampling followed by inspection of the neighboring vertices, in a time sub-linear in the network size.

In the context of online social networks, exploration often focused on how to discover the entire network more efficiently. Until recently this was feasible for many real world networks, before they exploded to their current size. It is no longer feasible to get a consistent snapshot of the Facebook network for example. (According to the Facebook statistics page at `www.facebook.com/press/info.php?statistics`, retrieved on 2 June 2011, there were over 500 million active users, and around 36 billion links.)

Methods based on random walks are commonly used for graph searching and crawling. Stutzbach *et al* [14] compare the performance of breadth first search (BFS) with a simple random walk and a Metropolis Hastings random walk on various classes of random graphs as a basis for sampling the degree distribution of the underlying networks. The purpose of the investigation was to sample from dynamic Peer-To-Peer (P2P) networks. In a related study Gjoka *et al* [11] made extensive use of the these methods to collect a sample of Facebook users. As simple random walks are degree biassed they used a re-weighting technique to unbias the sampled degree sequence output by the random walk. This is referred to as a re-weighted random walk in [11]. In both the above cases it was shown

the bias could be removed dynamically by using a suitable Metropolis-Hastings random walk.

A simple way to generate a graph with a power law degree sequence is to use the preferential attachment method described by Albert and Barabási [4]. In this model, the graph $G(t) = G(m, t)$ is obtained from $G(t-1)$ by adding a new vertex v_t with m edges between v_t and $G(t-1)$. The end points of these edges are chosen preferentially, that is to say proportional to the existing degree of vertices in $G(t-1)$. Thus the probability $p(x, t)$ that vertex $x \in G(t-1)$ is chosen as the end point of a given edge is equal to $p(x, t) = d(x, t-1)/(2m(t-1))$, and this choice is made independently for each of the m edges added. A model generated in this way has a power law of $c = 3$ for the degree sequence, irrespective of the number of edges $m \geq 1$ added at each step. For a graph constructed in this way, the expected degree at step t of the vertex added at step s is $\mathbf{E}d(s, t) \sim m(t/s)^{1/2}$.

The preferential attachment model was refined by Bollobas et al [5] who introduced the scale free model to make detailed calculations of degree sequence. The model was generalized by many authors, including the web-graph model of Cooper and Frieze [8]. The web-graph model is more general, and allows the number of edges added at each step to vary, for edges from new vertices to choose their end points preferentially or uniformly at random, and for insertion of edges between existing vertices. By varying these parameters, preferential attachment graphs with degree sequences exhibiting power laws c in the interval $(2, \infty)$ are obtained.

The power law c for preferential attachment graphs and web-graphs can be written explicitly as

$$c = 1 + 1/\eta, \tag{1}$$

where η is the expected proportion of edge end points added preferentially (see [7]). For example in the standard preferential attachment process (e.g. the Barabási and Albert model), $\eta = 1/2$, as each new edge chooses an existing neighbour vertex preferentially; thus explaining the power law of 3 for this model.

In the simplest case, to form $G(t)$, a new vertex v_t is added at each step t with m edges directed towards the existing graph $G(t)$. Each edge chooses its terminal vertex either by preferential attachment or uniformly at random with some probability mixture p or $1 - p$. This generates a power law $c \geq 3$. We refer to this generalized process as $G(c, m, t)$. For this example, the parameter $\eta = p/2$ depends on the proportion p of edge end points chosen preferentially (as opposed to uniformly at random). The parameter η in (1) occurs in process models, in the expression for the expected degree of a vertex. Let $d(s, t)$ denote the degree at step t of the vertex v_s added at step s. The expected value of $d(s, t)$ is given by

$$\mathbf{E}d(s, t) \sim m \left(\frac{t}{s} \right)^{\eta}. \tag{2}$$

Thus, in the preferential attachment model of [4], $\mathbf{E}d(s, t) \sim m(t/s)^{1/2}$.

Generalizing this, we consider arbitrary multi-graphs $G(t)$ on t vertices which, have the following properties, which we call *pseudo-preferential*.
(i) When the vertices are relabeled $s = 1, ..., t$ by sorting on vertex degree in descending order, $G(t)$ has a degree sequence which satisfies

$$\left(\frac{t}{s}\right)^{\eta(1-\epsilon)} \leq d(s) \leq \left(\frac{t}{s}\right)^{\eta} \log^2 t, \tag{3}$$

for some $\epsilon > 0$ and $0 < \eta < 1$, and for some range of $s \geq 1$.
(ii) For all vertices s in the sorted order, s has at most m edges to vertices $\sigma \leq s$.

Our particular aim is, given $a > 0$, to find all vertices $v \in V(t)$ of degree $d(v) \geq t^a$. Denote by S_a the set of vertices of $G(t)$ of degree $d(v) \geq t^a$. For the following reason, we will assume $a < \eta$. The maximum degree in (3) is $\widetilde{O}(t^{\eta})$, and, from (2), this is also the maximum expected degree in preferential attachment graphs ($\eta = 1/2$) and web-graphs ($0 < \eta < 1$). We use the notation $\widetilde{O}(f(t))$ as shorthand for $O(f(t)\log^k t)$ where t is the size of $G(t)$ and k is a positive constant.

We say a random walk is *seeded* if the walk starts from some vertex s of S. In the context of searching networks such as Facebook, Twitter or the WWW it is not unreasonable to suppose we know *some* high degree vertex without supposing we know all of them. Experimentally, we found the *seeding* condition was not necessary, but a general analysis without this condition would require notions of mixing time and stationarity which our analysis avoids. The following theorem holds for any network with the pseudo-preferential properties given above.

Theorem 1. *Let $G(t)$ be a pseudo-preferential graph with degree sequence satisfying (3). Let $S_a = \{v : d(v) \geq t^a\}$ be connected with diameter $Diam(S_a)$. Let $b = (1-\eta)/(\eta(2-\eta(1-\epsilon))$.*

*A biassed seeded random walk with transition probability along edge $\{x, y\}$ proportional to $(d(x)d(y))^b$, finds all vertices in $G(t)$ of degree at least t^a in $\widetilde{O}(Diam(S_a) \times t^{1-2ab(1-\epsilon)})$ steps, With High Probability (**whp**).*
The cover time of the graph $G(t)$ by this biassed walk is $\widetilde{O}(tDiam(G(t)))$.

In reality the degree sequence (3) of graph $G(t)$ is unknown, but η can be estimated as $\eta = 1/(c-1)$ from the power law c of the degree sequence. Optimistically setting $\epsilon = 0$ then gives a value b for the search algorithm. Its also fair to say that, experimentally, we found putting $b = 1/2$ in the biassed random walk was effective a variety of real networks with a power law degree sequence.

We next give a general result for web-graphs $G(c, m, t)$, which is also valid for related models such as scale free graphs. For the class of graphs $G(c, m, t)$, the lower bound on the degree of vertex s becomes less concentrated as s tends to t, so that the value of ϵ we must choose for our lower bound in (3) increases with

s. Thus, as the vertex degree t^a decreases, the upper bound on the algorithm runtime increases in a way which depends on a, c, m. As long as we incorporate this dependence, Theorem 2 says that if we search $G(c, m, t)$ using a random walk with a bias b proportional to the power law c then, (i) we can find all high degree vertices quickly, and (ii) the time to discover all vertices is of about the same order as for a simple random walk.

Theorem 2. *Let $c \geq 3$, and let $m \geq 2$. Let $a < 1/(c-1)$, and let $\epsilon = (1+1/a-c)/m$.*

*Let $b = (c-1)(c-2)/(2c-3)$. For $c \geq 3$, **whp** we can find all vertices in $G(c, m, t)$ of degree at least t^a in $\widetilde{O}(t^{1-2ab(1-\epsilon)})$ steps, using a biassed seeded random walk with transition probability along edge $\{x, y\}$ proportional to $(d(x)d(y))^b$.*
The cover time of the graph $G(c, m, t)$ by this biassed walk is $\widetilde{O}(t)$.

The maximum degree of $G(c, m, t)$ is $\widetilde{O}(t^\eta)$ **whp**, where $\eta = 1/(c-1)$ which explains the bound on a given above. Using this, a t^{1-2ab} run time can be repackaged as follows. Let $a = \theta\eta$ for $0 < \theta < 1$, then $2ab = \theta(1 - 1/(2c-3))$.

2 Properties of the web-graph process

The actual value of $d(s, t)$ is not concentrated around $\mathbf{E}d(s, t)$ in the lower tail, but the following inequality is adequate for our proof.

Lemma 1. *Given $G(c, m, t)$ and a, ϵ and suppose $m > (1/\epsilon)(1/a - 1/\eta)$.*

With high probability for all vertices s, such that $\mathbf{E}d(s, t) \geq t^a$, we have that $d(s, t) \geq \left(\frac{t}{s}\right)^{\eta(1-\epsilon)}$. For all $s \geq \log^2 t$, $d(s, t) \leq \left(\frac{t}{s}\right)^{\eta} \log^2 t$.

For proof of Lemma 1 see Appendix. We also need lower tail concentration for large sets of vertices.

Lemma 2. *Let $d([s], t)$ denote the total degree at step $t \geq s$ of the set $[s] = \{1, ..., s\}$. Let $K > 1$. Then*

$$\mathbf{Pr}\left(d([s], t) \leq \frac{2ms}{K}\left(\frac{t}{s}\right)^{\eta}\right) = O(s^{-mK}).$$

The upshot of this, is that all vertices added after step $v = s\log^{2/\eta+1} t$ have degree $d(v, t) = o\left((t/s)^\eta\right)$ **whp**. This observation forms the basis of our sub-linear algorithm. For proof of Lemma 2 see Appendix.

Another piece of the puzzle we will need, is that **whp** web-graphs have diameter

$$\mathrm{Diam}(G(c,m,t)) = O(\log t) \tag{4}$$

Crude proofs of this can be made for the web-graph model based on expansion properties of the graph. For example, in the preferential attachment graph ($\eta = 1/2, c = 3$) when vertex t is added to $G(m,t)$ the probability that t does not select at least one neighbour in $G(t/2)$ is at most

$$\left(1 - \frac{2m(t/2)}{2mt}\right)^m = \left(\frac{1}{2}\right)^m.$$

Thus $\mathrm{Diam}(G(m,t)) = O(\log t)$ by a 'tracing backwards stochastically' argument.

3 Biassed random walks

Let $G = (V,E)$ be a connected undirected graph. A *random walk* W_u, $u \in V$, on G is a Markov chain $X_0 = u, X_1, \ldots, X_t, \ldots$ on the vertices V associated to a particle that moves from vertex to vertex according to a transition rule. The probability of a transition from vertex i to vertex j is $p(i,j)$ if $\{i,j\} \in E$, and 0 otherwise.

Let $d(v) = d(v,t)$ be the degree of vertex $v \in G(t)$, and let $N(v)$ denote the neighbours of v in this graph. The basis of our algorithm is a degree-biassed random walk, with transition probability $p(u,v)$ given by

$$p(u,v) = \frac{(d(v))^b}{\sum_{w \in N(u)} (d(w))^b}, \tag{5}$$

where $b > 0$ constant. The value of b we will choose in our proof is optimized to depend on η. Thus for Theorem 2, using (1), the value of b can be expressed directly as a function of the degree sequence power law c.

The easiest way to reason about biassed random walks, is to give each edge e a weight $w(e)$, so that transitions along edges are made proportional to this weight. In the case above the weight of the edge $e = (u,v)$ is given by $w(e) = (d(u)d(v))^b$ so that the transition probability (5) is now written as

$$p(u,v) = \frac{(d(u)d(v))^b}{\sum_{w \in N(u)} (d(u)d(w))^b}. \tag{6}$$

The inspiration for the degree biassed walk above, comes from the β-walks of Ikeda, Kubo, Okumoto and Yamashita [12] which use an edge weight $w(x,y) =$

$1/(d(x)d(y))^\beta$ to favor low degree vertices. When $\beta = 1/2$ this gives an improved worst case bound of $O(n^2 \log n)$ for the cover time of connected n-vertex graphs.

We next note some facts about weighted random walks, which can be found in Aldous and Fill [2] or Lovasz [13]. The weight $w(e)$ of an edge e has the meaning of conductance in electrical networks, and the resistance $r(e)$ of e is given by $r(e) = 1/w(e)$. The commute time $K(u, v)$ between vertices u and v, is the expected number of steps taken to travel from u to v and back to u. The commute time for a weighted walk is given by

$$K(u, v) = w(G)R_{\mathit{eff}}(u, v). \tag{7}$$

Here $w(G) = 2\sum_{e\in E} w(e)$ and $R_{\mathit{eff}}(u, v)$ is the effective resistance between u and v, when G is taken as an electrical network with edge e having resistance $r(e)$. For our proof we do not need to calculate $R_{\mathit{eff}}(u, v)$ very precisely, but rather note that if uPv is any path between u and v then

$$R_{\mathit{eff}}(u, v) \le \sum_{e\in uPv} r(e).$$

For $u \in V$, and a subset of vertices $S \subseteq V$, let $C_u(S)$ be the expected time taken for W_u to visit every vertex of G. The *cover time* C_S of S is defined as $C_S = \max_{u\in V} C_u(S)$. We define a walk as *seeded* if it starts in S. The *seeded cover time* ${C_S}^*$ of S as ${C_S}^* = \max_{u\in S} C_u(S)$. For a random walk starting in a set S, the cover time of S satisfies the following Matthews bound

$$C^*_S \le \max_{u,v\in S} H(u, v) \log |S|. \tag{8}$$

For $u \neq v$, the variable $H(u, v)$ is the expected time to reach v starting from u (the hitting time). The commute time $K(u, v)$ is given by $K(u, v) = H(u, v) + H(v, u)$, so $K(u, v) > H(u, v)$.

4 Proof of Theorems 1 and 2

We apply the Matthews bound (8). Clearly $\log |S_a| \le \log t$. It remains to find

$$\max_{u,v\in S} H(u, v) \le \max_{u,v\in S} K(u, v).$$

To calculate $K(u, v)$ in (7), we first need to bound $w(G)$

Lemma 3. *By choosing*

$$b = \frac{1-\eta}{\eta(2-\eta(1-\epsilon'))},$$

where $\epsilon' = \epsilon$ for Theorem 1, and $\epsilon' = 0$ for Theorem 2, it follows that $w(G) = O(t \log^{4b+1} t) = \widetilde{O}(t)$

Proof

We define a graph G^* on vertices $1, 2, \ldots, t$ which has the same degree sequence as graph G, and is built in a similar iterative process: for each $v = t_0, t_0 + 1, \ldots, t$, add m edges from vertex v to some earlier vertices. In graph G, edges are selected according to a random preferential process, while in graph G^* according to the deterministic process which greedily fills the in-degrees of vertices, giving preference to the older vertices. In both graphs, if (x, y) is a directed edge, then $y < x$ (the edges point from x towards the earlier vertex y).

Assume $b > 0$ and define

$$\bar{d}(v) = \left(\frac{t}{v}\right)^{\eta},$$
$$\bar{w}(G) = 2 \sum_{\{x,y\} \in E(G)} \left(\bar{d}(x)\bar{d}(y)\right)^b \;\geq\; w(G) \log^{-4b} t.$$

Graph G^* is obtained from G by repeatedly swapping edges. Whenever there is a pair of edges (x, y), (u, v) such that $x < u$ but $y > v$, then replace them with edges (x, v) and (u, y). If $A > B$ and $C > D$ then $(A - B)(C - D) > 0$ so $AC + BD > AD + BC$. Thus each swap increases $\bar{w}(G)$ because

$$(\bar{d}(x))^b > (\bar{d}(u))^b \quad \text{and} \quad (\bar{d}(y))^b < (\bar{d}(v))^b$$

implies

$$(\bar{d}(x))^b(\bar{d}(v))^b + (\bar{d}(u))^b(\bar{d}(y))^b \;>\; (\bar{d}(x))^b(\bar{d}(y))^b + (\bar{d}(u))^b(\bar{d}(v))^b.$$

Therefore, $\bar{w}(G^*) \geq \bar{w}(G)$. By construction, a vertex v in G^* has incoming edges originating from vertices $\text{first}(v), \text{first}(v) + 1, \ldots, \text{last}(v)$. Thus we have

$$\begin{aligned}
\bar{w}(G^*) &= 2 \sum_{\{y,x\} \in E(G^*)} \left(\bar{d}(x)\bar{d}(y)\right)^b \\
&= 2 \sum_{x=1}^{t} \sum_{y=\text{first}(x)}^{\text{last}(x)} \left(\bar{d}(x)\bar{d}(y)\right)^b \\
&\leq 2 \sum_{x=1}^{t} d(x) \left(\bar{d}(x)\bar{d}(\text{first}(x))\right)^b \\
&\leq 2 \sum_{x=1}^{t} \left(\bar{d}(x)\right)^{1+b} \left(\bar{d}(\text{first}(x))\right)^b . \qquad (9)
\end{aligned}$$

Now we calculate $\text{first}(x)$. The $m(\text{first}(x) - 1)$ edges outgoing from vertices $1, 2, \ldots, \text{first}(x) - 1$ fully fill the in-degrees of vertices $1, 2, \ldots, x - 1$, so

$$m \cdot \text{first}(x) = 1 + \sum_{z=1}^{x-1} (d(z) - m).$$

Let C be some generic constant whose value can vary. For Theorem 1 choosing $\epsilon' = \epsilon$, (deterministic case),

$$\sum_{z=1}^{x-1} d(z) \geq \sum_{z=1}^{x-1} \left(\frac{t}{z}\right)^{\eta(1-\epsilon)} = Ct^{\eta(1-\epsilon')} x^{1-\eta(1-\epsilon')}.$$

For Theorem 2 (web-graph case), choosing $\epsilon' = 0$ we have from Lemma 2 that

$$\sum_{z=1}^{x-1} d(z) \geq mx \left(\frac{t}{x}\right)^{\eta}.$$

Thus

$$\bar{d}(\text{first}(x)) \leq C \left(\frac{t}{t^{\eta(1-\epsilon')} x^{1-\eta(1-\epsilon')}}\right)^{\eta} = C \left(\frac{t}{x}\right)^{\eta(1-\eta(1-\epsilon'))}. \tag{10}$$

Using (10) in (9), we get

$$\begin{aligned} \bar{w}(G^*) &\leq C \sum_{x=1}^{t} \left(\frac{t}{x}\right)^{\eta(1+b)} \left(\frac{t}{x}\right)^{b\eta(1-\eta(1-\epsilon'))} \\ &= C \sum_{x=1}^{t} \left(\frac{t}{x}\right)^{\eta(1+b(2-\eta(1-\epsilon')))}. \end{aligned} \tag{11}$$

Choosing

$$\eta(1 + b(2 - \eta(1 - \epsilon'))) = 1, \tag{12}$$

the sum in (11) is $O(t \log t)$ and we have

$$w(G) \leq \log^{4b} \bar{w}(G) \leq \log^{4b} \bar{w}(G^*) = O(t \log^{4b+1} t). \tag{13}$$

□

Details specific to Theorem 1. The set S_a is connected with diameter $\text{Diam}(S_a)$ is as specified. Let $\Delta(a) = \text{Diam}(S_a)$, then for any $u, v \in S_a$ there is a path uPv of length $O(\Delta(a))$ from u to v in $G(t)$ contained in S_a, and thus consisting of vertices w of degree $d(w,t) \geq (t/s)^{\eta(1-\epsilon)} = d^*$. Thus all edges of this path have resistance at most $1/(d(x)d(y))^b \leq 1/(d^*)^{2b}$.

Details specific to Theorem 2. Suppose we want to find all vertices of degree at least t^a for some $a > 0$ in $G(t) \equiv G(c, m, t)$. Let $S_a = \{v : d(v,t) \geq t^a\}$. Recall that $G(t)$ is generated by a process of attaching v_t to $G(t-1)$. At what steps were the vertices $v \in S_a$ added to $G(t)$? The expected degree of v at step t is given by (2) i.e. $\mathbf{E}d(v,t) = (1 + o(1))m(t/v)^{\eta}$. This function is monotone decreasing with increasing v. Let σ be given by

$$t^a = \left(\frac{t}{\sigma}\right)^{\eta} \qquad \text{which implies} \qquad \sigma = t^{1-a/\eta}. \tag{14}$$

Let $s = \sigma \cdot \log^{2/\eta+1} t$, then using (3) all vertices added at steps $w \geq s$ have $d(w,t) = o(t^a)$. On the other hand, using (3) again, all vertices v added at steps $1, ..., s$ have degree $d(v,t) \geq (t/s)^{\eta(1-\epsilon)}$.

For Theorem 2 let $\Delta(a) = \mathrm{Diam}(G(s))$ where s is as defined above. Because $\mathrm{Diam}(G(s)) = O(\log s)$, (see (4)), we know that for any $u, v \in S_a$ there is a path uPv of length $O(\log t)$ from u to v in $G(t)$ contained in $G(s)$, and thus consisting of vertices w of degree $d(w,t) \geq (t/s)^{\eta(1-\epsilon)} = d^*$. Thus all edges of this path have resistance at most $1/(d(x)d(y))^b \leq 1/(d^*)^{2b}$.

Proof of Theorems 1 and 2. From (3), d^* satisfies

$$d^* \geq \left(\frac{t}{t^{1-a/\eta} \log^{1+2/\eta} t} \right)^{\eta(1-\epsilon)} \geq \frac{t^{a(1-\epsilon)}}{\log^3 t}.$$

By the discussion above,

$$R_{\mathit{eff}}(u,v) \leq \sum_{e \in uPv} r(e) = O\left(\frac{\Delta(a)}{d^*} \right).$$

Using (7), and the value of d^*, we have

$$K(u,v) \leq K^* = \widetilde{O}(\Delta(a) t^{1-2ba(1-\epsilon)}).$$

The bound in Theorem 2 on finding all vertices of degree at least t^a is now obtained as follows. The Matthews bound (8) gives the (expected) cover time $C^*_{S_a} = O(K^* \log t)$. Apply the Markov inequality ($\mathbf{Pr}(X > A \cdot \mathbf{E}X) \leq 1/A$), with $\mathbf{E}X = C^*_{S_a}$, and $A = \log t$ to give a **whp** result, that all vertices of degree at least t^a can be found in time

$$T(a) = \widetilde{O}(t^{1-2ba(1-\epsilon)}).$$

For preferential attachment graphs $\eta = 1/2$, and (12) gives $b = 2/3$, and the time $T(a)$ is

$$\widetilde{O}(t^{1-(4/3)a(1-\epsilon)}).$$

Finally we establish the cover time of the graph $G(t)$. This is done by using (8) with $S = V(t)$ the vertex set of $G(t)$, i.e.

$$C_{V(t)} \leq \max_{u,v \in V(t)} H(u,v) \log t. \tag{15}$$

We bound $H(u,v)$ by (7) as usual. The resistance $r(e)$ of any edge $e = \{x, y\}$ is

$$r(e) = \frac{1}{(d(x)d(y))^b} \leq \frac{1}{m^{2b}} = O(1).$$

Let the diameter of $G(t)$ be $\mathrm{Diam}(G)$ which is specified for Theorem 1 and is $O(\log t)$ (**whp**) for Theorem 2. Thus $R_{\mathit{eff}}(u,v) = O(\mathrm{Diam}(G))$, since the effective resistance between u and v is at most the resistance of a shortest path between u and v. This and (13) give $K(u,v) = \widetilde{O}(t\mathrm{Diam}(G))$. Thus the cover time of the graph $G(t)$ is $\widetilde{O}(t\mathrm{Diam}(G))$.

5 Experimental results

Theorem 2 gives an encouraging upper bound of the order of around $t^{1-(4/3)a}$ for a biassed random walk to the cover all vertices of degree at least t^a in the t-vertex preferential attachment graph $G(3, m, t)$. Our experiments, summarized in Figure 2, suggest that the actual bound is stronger than this. The experiments were made on $G(m, t)$ with $m = 4$, and $t = 5 \times 10^6$ vertices. The degree distribution of this graphs is given in Figure 1, with both axes in logarithmic scale. More precisely, the x-axis is the exponent a in the degree $d = t^a$, i.e. $x = \log d / \log t$, while the y-axis is the frequency of the vertices of degree t^a.

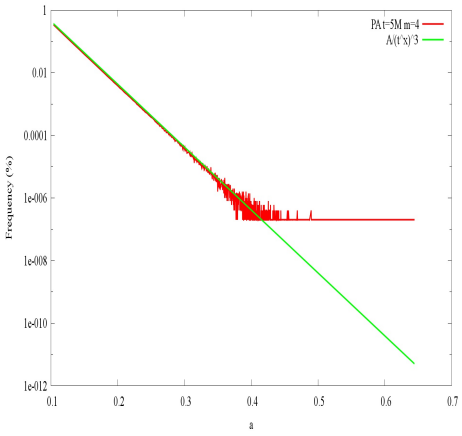

Fig. 1. Degree distribution of a realization of $G(c, m, t)$, $c = 3, m = 4, t = 5 \times 10^6$

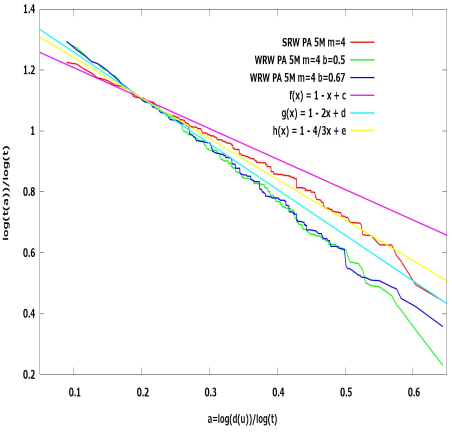

Fig. 2. Cover time of vertices of degree at least t^a in $G(3, 4, 5 \times 10^6)$ as a function of a.

In Figure 2, plot SRW shows the average cover time $\tau(a)$ of all vertices of degree at least t^a by the simple random walk (the uniform transition probabilities). Plot WRW shows the average cover times by the biassed random walk with $b = 1/2$, and $b = 2/3$. The plots are an average of 9 runs (each) of the random walks. Both axes are in logarithmic scale. The y-axis is $y = (\log \tau(a))/\log t$. There are also three reference lines drawn in Figure 2. These lines have slopes $-a$, $-4a/3$ and $-2a$, and are included for visual inspection only. To calculate the speed up, given $x = a$, read off the $y(a)$-values y_S, y_W. The speed up is $t^{y_S - y_W}$, where $t = 5 \times 10^6$. In the upper tail the weighted walks is about 10 times faster. Curiously, the improvement does not seem sensitive to the precise value of b.

The cover time C_G of a simple random walk on $G(m, t)$ is known and has value $C_G \sim (2m/(m-1))t \log t$, see [9]. The intercept of the y-axis predicted by this is $y_C = \log C_G / \log t$, which when $m = 4$ and $t = 5 \times 10^6$ is $y_C = 1.29$ This agrees

well with the experimental intercept of 1.24, and helps confirm the accuracy of our simulations.

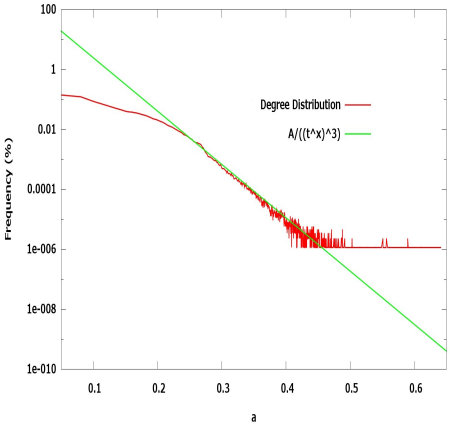

Fig. 3. Degree distribution of sample of size 8.7×10^5 of G_W, the underlying graph of the WWW

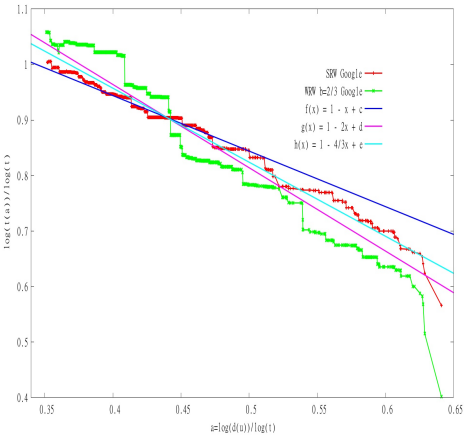

Fig. 4. Cover time of vertices of degree at least t^a in G_W as a function of a

Our experimental results for Theorem 1 are less clear cut, but still encouraging. Figure 3 gives the degree distribution of the underlying graph of the WWW, on $t = 8.7 \times 10^5$ vertices obtained from `http://snap.stanford.edu/data/web-Google.html`. The power law exponent is approximately $c = 3$, and it was crawled using a weight of $b = 2/3$. Figure 4, shows the results obtained by averaging 25 runs of the simple and weighted random walks. The weighted walk is generally about 4 times faster for $a > 0.43$.

References

1. D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *J. ACM*, 56(4), 2009.
2. D. Aldous and J. A. Fill. Reversible Markov chains and random walks on graphs. http://stat-www.berkeley.edu/pub/users/aldous/RWG/book.html, 1995.
3. R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country: Better strategies than breadth-first for web page ordering. In *Proc. 14th International Conference on World Wide Web*, pages 864–872. ACM Press, 2005.
4. A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, number 5439 in Volume 286, pages 509–512, 1999.
5. B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18:279–290, 2001.

6. M. Brautbar and M. Kearns. Local algorithms for finding interesting individuals in large networks. In *Proceedings of ICS 2010*, pages 188–199, 2010.
7. C. Cooper. The age specific degree distribution of web-graphs. *Combinatorics Probability and Computing*, 15:637–661, 2006.
8. C. Cooper and A. Frieze. A general model web graphs. In *Random Structures and Algorithms*, vol. 22(3), pages 311–335, 2003.
9. C. Cooper and A. Frieze. The cover time of the preferential attachment graphs. *Journal of Combinatorial Theory*, B(97):269–290, 2007.
10. A. D. Flaxman and J. Vera. Bias reduction in traceroute sampling - towards a more accurate map of the Internet. In *Proceedings of WAW 2007*, pages 1–15, 2007.
11. M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. A walk in Facebook: Uniform sampling of users in online social networks. *CoRR*, abs/0906.0060, 2009.
12. S. Ikeda, I. Kubo, N. Okumoto, and M. Yamashita. Impact of Local Topological Information on Random Walks on Finite Graphs. In *Proceedings of ICALP 2003*, pages 1054–1067.
13. L. Lovász. Random walks on graphs: A survey. *Bolyai Society Mathematical Studies*, 2:353–397, 1996.
14. D. Stutzbach, R. Rejaie, N.G. Duffield, S. Sen and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement - IMC 2006*, pages 27–40, 2006.

Appendix

Lemma 1 Given $G(c, m, t)$ and a, ϵ and suppose $m > (1/\epsilon)(1/a - 1/\eta)$.

With high probability for all vertices s, such that $\mathbf{E}d(s,t) \geq t^a$, we have that $d(s,t) \geq \left(\frac{t}{s}\right)^{\eta(1-\epsilon)}$. For all $s \geq \log^2 t$, $d(s,t) \leq \left(\frac{t}{s}\right)^{\eta} \log^2 t$.

Proof The upper bound on $d(s,t)$ is given in [7], as is the following degree distribution. For $m \geq 2$, the distribution of $d(s,t)$ is given by

$$\mathbf{Pr}(d(s,t) = m + \ell \mid d(s,s) = m) \leq C \binom{m+\ell-1}{\ell} \left(\frac{s}{t}\right)^{\eta m} \left(1 - \left(\frac{s}{t}\right)^{\eta}\right)^{\ell}.$$

Thus, crudely

$$\mathbf{Pr}(d(s,t) \leq \ell) \leq C\ell^m \left(\frac{s}{t}\right)^{\eta m}.$$

Inserting $\ell = \left(\frac{t}{s}\right)^{\eta(1-\epsilon)}$, and choosing $s = t^{1-a/\eta}$, we find the expected number of vertices $1 \leq v \leq s$ not satisfying the lower bound is of order

$$s \left(\frac{s}{t}\right)^{m\epsilon\eta} = t^{(1-a/\eta)(1+m\epsilon\eta)} = o(1),$$

provided

$$m > \frac{1}{\epsilon}\left(\frac{1}{a} - \frac{1}{\eta}\right).$$

□

Lemma 2 Let $d([s],t)$ denote degree of $[s] = \{1, ..., s\}$ at step t. Let $K > 1$. Then

$$\mathbf{Pr}\left(d([s],t) \leq \frac{2ms}{K}\left(\frac{t}{s}\right)^{\eta}\right) = O(s^{-mK}).$$

Proof We give the proof for $\eta = 1/2$ (preferential attachment), the general proof is similar.

Let $Z_t = d([s],t)$. Then $Z_t = X_t + Z_{t-1}$ where $Z_s = 2ms$ and $X_t \sim Bin(m, Z_{t-1}/(2m(t-1)))$. Also, $\mathbf{E}Z_t \sim 2ms(t/s)^{1/2}$. Given $h, c_t, A > 0$,

$$\mathbf{Pr}(Z_t < A) = \mathbf{Pr}(e^{-hZ_t/c_t} > e^{-hA/c_t}).$$

Let $p = Z_{t-1}/(2m(t-1))$, then

$$\begin{aligned}\mathbf{E}(e^{-hX_t/c_t}) =& (1 - p + pe^{-h/c_t})^m \\ \leq& e^{-\frac{h}{c_t}(1-h/c_t)\frac{Z_{t-1}}{2(t-1)}},\end{aligned}$$

by using $e^{-x} \leq 1 - x + x^2$. Let $c_s = 1, c_t = (1 + 1/(2(t-1)))c_{t-1}$ so that $c_t \sim (t/s)^{1/2}$. We will choose $h = o(1)$ (see below). Iterating the expression $Z_t = X_t + Z_{t-1}$, gives

$$\begin{aligned}\mathbf{E}(e^{-hZ_t/c_t}) \leq& e^{-h\frac{Z_{t-1}}{c_{t-1}}\frac{1+(1-h/c_t)/(2(t-1))}{1+1/(2(t-1))}} \\ =& e^{-h'Z_{t-1}/c_{t-1}},\end{aligned}$$

where

$$h(1 - O(h/tc_t)) \leq h' \leq h,$$

and

$$\mathbf{E}(e^{-hZ_s/c_s}) = e^{-h2ms}.$$

All in all,

$$\mathbf{E}(e^{-hZ_t/c_t}) \leq \mathbf{E}\left(e^{-h\frac{Z_s}{c_s}\prod_{j=s}^{t-1}(1-O(h/(jc_j)))}\right) = e^{-h2ms(1-O(h))}.$$

Choosing $A = \mathbf{E}Z_t/K'$ and applying the Markov inequality that $\mathbf{Pr}(Y \geq A) \leq \mathbf{E}(Y)/A$ with $Y = e^{-hZ_t/c_t}$, we have

$$\mathbf{Pr}(Z_t \leq \mathbf{E}Z_t/K') \leq e^{-h2ms(1-1/K'-O(h))} = O(s^{-mK}),$$

on choosing $h = (K \log s)/s = o(1)$. □