A Fast Algorithm to Find All High Degree Vertices in Power Law Graphs

Colin Cooper
Department of Informatics
King's College London
London, U.K.
colin.cooper@kcl.ac.uk

Tomasz Radzik
Department of Informatics
King's College London
London, U.K.
tomasz.radzik@kcl.ac.uk

Yiannis Siantos
Department of Informatics
King's College London
London, U.K.
yiannis.siantos@kcl.ac.uk

ABSTRACT

Sampling from large graphs is an area which is of great interest, particularly with the recent emergence of huge structures such as Online Social Networks. These often contain hundreds of millions of vertices and billions of edges. The large size of these networks makes it computationally expensive to obtain structural properties of the underlying graph by exhaustive search. If we can estimate these properties by taking small but representative samples from the network, then size is no longer a problem.

In this paper we develop an analysis of random walks, a commonly used method of sampling from networks. We present a method of biassing the random walk to acquire a complete sample of high degree vertices of social networks, or similar graphs. The preferential attachment model is a common method to generate graphs with a power law degree sequence. For this model, we prove that this sampling method is successful with high probability.

For t-vertex graphs $G(t)$ generated by a preferential attachment process, 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. Let $S(a)$ be the set of all vertices of degree at least t^a in $G(t)$. We show that for some $b \approx 2/3$, if the biassed random walk starts at an arbitrary vertex of $S(a)$, then with high probability the set $S(a)$ can be discovered completely in $\widetilde{O}(t^{1-(4/3)a+\delta})$ steps, where δ is a very small positive constant. The notation $\widetilde{O}$ ignores poly-log t factors.

The preferential attachment process generates graphs with power law 3, so the above example is a special case of this result. For graphs with degree sequence power law $c > 2$ generated by a generalized preferential attachment process, a random walk with transitions along undirected edges $\{x, y\}$ proportional to $(d(x)d(y))^{(c-2)/2}$, discovers the set $S(a)$ completely in $\widetilde{O}(t^{1-a(c-2)+\delta})$ steps with high probability. The cover time of the graph is $\widetilde{O}(t)$.

Our results say that if we search preferential attachment graphs with a bias $b = (c-2)/2$ 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 not much higher than in the case of a simple random walk. We conduct experimental tests on generated networks and real-world networks, which confirm these two properties.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

Categories and Subject Descriptors

H.3.3. [**Information Search And Retrieval**]: Search process; F.2.2. [**Nonnumerical Algorithms and Problems**]: Computations on discrete structures; G.2. [**Discrete Mathematics**]: Graph Theory—*Graph algorithms, Network Problems*; G.3. [**Probability and Statistics**]: Markov Processes

General Terms

Graph Theory, Random Walks

Keywords

graph sampling, random walks, weighted random walks

1. INTRODUCTION

The explosive growth of Online Social Networks (OSNs) over the past few years and the relative power that people have within these networks to affect the world around them, is a phenomena whose importance is beyond dispute. Recent developments in technology have allowed the creation of large networks, available globally via personal computers, or more recently mobile phones. The original and most outstanding example of such networks is the World Wide Web (WWW), and the email network. Very recently, many novel OSNs such as Twitter and Facebook, or online video repositories such as YouTube have sprung up. These networks, extensively interleaved with each other and the WWW, have substantial impact on the way our lives are lived. Nobody who was followed the political unrest in North Africa during February of 2011 can be unaware of the importance of Facebook and Twitter in galvanizing and coordinating social behavior.

The very large size of these networks makes it a major problem to get a good idea of their structure. This is especially true, given the limited amount of resources that are usually available when accessing them for research purposes. To solve this, a way to take small but representative samples of these networks needs to be found, which gives the correct idea of the structure of the entire graph.

While the intuition for modeling OSNs as a graph is rather simple, one may discover that in interpreting the correct structure, or proposing a generative model, even from large samples, there are many questions that automatically arise.

Some questions revolve around determining the best theoretical generation model to simulate such networks and others in discovering good methods of obtaining samples of small parts of the networks, which are representative and thus maintain the same properties as the entire network.

The area of graph sampling mentioned above is a developing topic, and many open questions exist. Finding effective and efficient ways to sample online graphs are of great interest, a useful tool for the community, and if successful should have practical applications.

The problem we consider, is how to find quickly all high degree vertices of a graph generated by preferential attachment. Although the distribution of node degrees in such graphs is heavy tailed and they have a larger number of high degree vertices than a random graph of equivalent edge density, the actual number of such vertices is small compared to the total number of vertices in the graph. Any standard sampling methods such as uniform sampling, simple random walks, fixed depth BFS etc would find it difficult to locate any, let alone all such vertices within a small sample.

2. RELATED WORK

2.1 Scale-Free and power-law graphs

The WWW graph in particular has received a great deal of attention. This graph is the result of modeling the WWW as a set of pages which are the vertices of the graph, and a set of directed links between the pages which are the edges of the graph [20]. There are some very interesting properties that have been discovered in this graph, for example the degree distribution observed follows a long-tail distribution [23] [9]. This is apparently a common attribute that is present in the WWW graph, as well as the Internet (autonomous systems) [16], citation graphs [25], OSNs [16] and many others.

The graphs with the above properties are commonly referred to as *power-law* graphs or *scale-free* graphs, however the latter term is controversial. For simplicity's sake we will use the term power-law graphs and scale-free graphs interchangeably to refer to all graphs which have power-law degree distributions, and a diameter less than or equal to the expected diameter of a small-world network. Additionally we will require our scale-free graphs to have a scale invariance where these basic properties remain true even at very different sizes of the graph. According to the work of Dill *et al*[14] the WWW graph exhibits a self-similar structure, which may be the cause of the aforementioned scale invariance of this graph. It may be reasonable to assume that many other scale-free graphs observed such as OSN graphs share this characteristic with the WWW graph.

2.2 Preferential attachment graphs

The preferential attachment model is a graph process used to generate graphs with degree distributions which follow a power-law. This process was proposed by Barabási and Albert [4] as a generative procedure for a model of the WWW. Surveys by Bollobás and Riordan [5] and Drinea, Enachescu and Mitzenmacher [15] give many related generative procedures to obtain graphs with power-law degree sequences. The general idea, is to begin with an initial non-empty graph containing at least one edge. During each step, a new vertex is added to the network and is then connected to a number of existing vertices chosen according to probabilities proportional to their degree.

2.3 Graph Crawling and Random walks

Generally speaking, graph crawling is a very underdeveloped topic. Put simply, the big question is: 'How can one sample only a part of a graph and yet retain certain structural information that is present on the entire graph'. This question has many interpretations and shades of meaning. In our case, for example, we are interested in getting a crawled sample of a preferential attachment graph which is a good model of the graph in its entirety. Thus, this sample studied on its own will have certain required properties which need to hold for us. In particular we want the degree distribution that is observed in the entire network to be, at scale, observed in our crawled sample of the network. Other typical examples might be clustering coefficient or diameter.

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], [17]. Another approach, analysed in [8], 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, and such methods have been used and analyzed extensively. Stutzbach *et al* [26] 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* [18] made extensive use of the above 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 [18]. In both the above cases it was shown the bias could be removed dynamically by using a suitable Metropolis-Hastings random walk. This indicates that there are application or network specific optimizations that can be done on random walks in order to tune them to the required task.

An interesting experimental analysis on sampling methods such as Respondent Driven Sampling (RDS) and Metropolis-Hastings Random Walk has been done by Rasti *et al* [24], which shows the effect of graph structure and size on the efficiency of these methods. Several graph types were used in [24], including the Erdos-Renyi random graph, the Small World graph, the Barbasi-Albert (preferential attachment)

graph and the Hierarchical Scale-Free graph (a scale-free graph which has a structure of clusters within clusters). It was shown that the above sampling methods had a reduced efficiency when applied to the Hierarchical Scale-Free graph.

In a related study by Leskovec *et al* in [21] it is mentioned that there are two distinct goals in sampling a network: the **back in time** goal, where we would be interested in a sample which would look like a snapshot of the graph in an earlier epoch or time period, and the **scale-down** goal, where we would be interested in a sample of the graph in which properties of the current state of the graph are preserved in proportion to the sample size.

In most cases we are just interested in a very specific properties of the graph. In that respect scaled down sampling of the graph may be focused primarily around obtaining these properties. For example, if we need to know the degree distribution of the network, it may be sufficient to get a subset of vertices sampled uniformly at random, and examine the proportion of vertices of each degree appearing in the sample. Even in this simple case however, different properties of the degree sequence may need several distinct sampling processes in order to get a good overall picture. For example, a sample with the same average degree as the graph may give no information about high degree vertices. Obtaining information about average distance between vertices might require a completely different process.

Our work differs from the aforementioned work with respect to the sampling goal. We wish to obtain all high degree vertices which are rare but significant vertices in such networks and therefore are hard to sample but important to have obtained.

3. OUR CONTRIBUTIONS

Preferential attachment graphs 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 features 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.

In our work we will focus on sampling the higher degree vertices, both the middle range and upper tail. Our aim is to sample *all these vertices*, and we propose a provably efficient method of obtaining those vertices in sub-linear time using a weighted random walk. Our reason for sampling all the higher degree vertices is that the upper tail is not concentrated, so no 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.

The simplest way to generate a graph with a power law degree sequence is 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 3 for the degree sequence, irrespective of the number of edges $m \geq 1$ added at each step.

Let S be a subset of the vertices of a graph $G = (V,E)$, where S is defined in terms of some property, such as the set of vertices with degree at least d. We suppose the content of S is unknown, and that we wish to discover all vertices in S by searching G using a random walk. 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.

THEOREM 1. *Let $G(m,t)$ be a graph generated in the preferential attachment model. Then With High Probability (**whp**) we can find all vertices in $G(m,t)$ of degree at least t^a in $O(t^{1-(4/3)a(1-\delta)})$ steps, using a biassed seeded random walk with transition probability along edge $\{x,y\}$ proportional to $(d(x)d(y))^{2/3}$. Here $\delta > 0$ is a small positive constant (eg. $\delta = 0.00001$). The cover time of $G(m,t)$ by this biassed walk is $O(t\,polylog\,t)$.*

The preferential attachment model was refined by Bollobas and Riordan [7], [6] who introduced the scale free model to make detailed calculations of degree sequence and diameter. The model was generalized by many authors, including the web-graph model of Cooper and Frieze [12], whose results we will need in our proofs below. The web-graph model is very 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, as well as 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. Assuming that m edges are added at every step, we refer to this generalized (web-graph) process with power law c as $G(c,m,t)$. Our motivation for considering this generalized process is to extend our analysis to networks whose power laws have been determined experimentally to be $c > 2$, but $c \neq 3$.

THEOREM 2. *For $c > 2$, **whp** we can find all vertices in $G(c,m,t)$ of degree at least t^a in $O(t^{1-a(c-2)(1-\delta)})$ steps, using a biassed seeded random walk with transition probability along edge $\{x,y\}$ proportional to $(d(x)d(y))^{(c-2)/2}$. Here $\delta > 0$ is a small positive constant (eg. $\delta = 0.00001$). The cover time of $G(c,m,t)$ by this biassed walk is $O(t\,polylog\,t)$.*

Theorem 2 says that if we search this type of graph using a random walk with a bias $b = (c-2)/2$ 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 a simple random walk. Theorem 1 gives a stronger bound for the special case of the (pure) preferential attachment model. We also conducted experimental tests on the preferential attachment model, which confirm these properties of the biassed random walk.

4. VERTEX DEGREES AND DIAMETER OF THE WEB-GRAPH PROCESS

In [11], Cooper noted the result that 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. In 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.

The value η occurs naturally in such 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

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

where η is the parameter defined above (see e.g. [10]). Thus, in the preferential attachment model of [4], $\mathbf{E}d(s,t) \sim m(t/s)^{1/2}$.

The actual value of $d(s,t)$ is not particularly concentrated around $\mathbf{E}d(s,t)$, but the following inequalities proved in e.g. [10] and [11], are adequate for our proofs. The inequalities hold with high probability (**whp**), for all vertices in $G(c,m,t)$.

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

where $\epsilon > 0$ is some arbitrarily small positive constant (e.g. $\epsilon = 0.00001$). The upshot of this, and our reason for explaining this to the reader, is that all vertices v added after step $s\log^{2/\eta+1} t$ have degree $d(v,t) = o\left((t/s)^{\eta}\right)$ **whp**. This observation forms the basis of our sub-linear algorithm.

The final piece of the puzzle we will need, is that the generalized web-graphs have diameter

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

with high probability. This was improved for scale free graphs by Bollobas and Riordan, but crude proofs can be made for the general web-graph model based on expansion properties of the graph.

5. 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 $b = (1/\eta - 1)/2$ we will choose in the proof of Theorem 2 below is optimized to depend on η. Using (1), this value can be expressed directly as a function of the degree sequence power law c, giving $b = (c-2)/2$.

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 a degree biassed walk with parameter b comes from the β-walks of Ikeda, Kubo, Okumoto and Yamashita [19] which use an edge weight $w(x,y) = 1/(d(x)d(y))^{\beta}$. 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 random walks, which can be found either in Aldous and Fill [2] or Lovasz [22]. 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 general theory of weighted random walks is given in Chapter 3 of [2].

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 $u \ne v$, 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 $K(u,v) > H(u,v)$.

6. PROOF OF 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 $S(a) = \{v : d(v,t) \ge t^a\}$. 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

$$t^a = \left(\frac{t}{\sigma}\right)^{\eta} \qquad \text{which implies} \qquad \sigma = t^{1-a/\eta}. \tag{9}$$

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 $d(v,t) \geq (t/s)^{\eta(1-\epsilon)}$.

We want to apply the Matthews bound (8). Clearly $\log |S(a)| \leq \log t$. It remains to find

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

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

$$\begin{aligned} w(G) &= 2 \sum_{\{x,y\} \in E(G)} w(x,y) \\ &= \sum_{x \in V} \sum_{y \in N(x)} (d(x)d(y))^b \\ &\leq \sum_{x \in V} \sum_{y \in N(x)} \frac{(d(x))^{2b} + (d(y))^{2b}}{2} \\ &= \sum_{x \in V} (d(x))^{2b+1} \\ &\leq \sum_{x=1}^{t} \left(\frac{t}{x}\right)^{\eta(2b+1)} \log^4 t. \end{aligned}$$

The upper bound on vertex degree in the last line comes from (3). Thus on choosing $\eta(2b+1) = 1$, that is, for $b = (1/eta - 1)/2$, we have

$$w(G) = O(t \log^5 t). \tag{10}$$

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 $1/(d(x)d(y))^b \leq 1/(d^*)^{2b}$. 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,

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

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

$$K(u,v) \leq K^* = O(t^{1-2ba(1-\epsilon)} \log^{12} t).$$

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 $C^*_{S(a)} = O(K^* \log t)$. Let $\delta = 3\epsilon$, an arbitrary but small constant. We use one of the ϵ to absorb the polylog term in K^*, and the other to apply the Markov inequality ($\mathbf{Pr}(X > A \cdot \mathbf{E}X) \leq 1/A$), with $\mathbf{E}X = C^*_{S(a)}$, to give a **whp** result.

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.

$$C_{V(t)} \leq \max_{u,v \in V(t)} H(u,v) \log t. \tag{11}$$

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

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

From (4) the diameter of $G(t)$ is $O(\log t)$, so $R_{\mathit{eff}}(u,v) = O(\log t)$, since the effective resistance between u and v is at most the resistance of a shortest path between u and v. This and (10) give $K(u,v) = O(t \log^6 t)$. Thus the cover time of the graph $G(t)$ is $O(t \log^7 t)$.

7. PROOF OF THEOREM 1

We consider now the preferential attachment graph $G(t) \equiv G(m,t)$. In this special case, $\eta = 1/2$ and the **whp** bounds (3) on the degree of vertex s are

$$\left(\frac{t}{s}\right)^{(1-\epsilon)/2} \leq d(s,t) \leq \left(\frac{t}{s}\right)^{1/2} \log^2 t. \tag{12}$$

We define a graph G^* on vertices $1, 2, \ldots, t$, which has the same degrees of vertices as in graph $G(t)$, and is built in a similar iterative process: for each $v = t_0 + 1, \ldots, t$, add m edges from vertex v to some earlier vertices. Graph $G(t_0)$ is the same constant-size starting graph for both $G(t)$ and G^*. In graph $G(t)$, 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 the preference to the older vertices. In both graphs, if $\{x,y\}$ is an edge and $x > y$, then this edge was added to the graph when vertex x was considered. Graph G^* can be obtained from graph $G(t)$ by swapping edges: whenever there is a pair of edges $\{x,y\}$ and $\{u,v\}$ such that $u > x > y > v$, then replace these edges with edges $\{x,v\}$ and $\{u,y\}$.

Assume $b > 0$ and define

$$\begin{aligned} \bar{d}(v) &= \left(\frac{t}{v}\right)^{1/2}, \\ \bar{w}(G) &= 2 \sum_{\{x,y\} \in E(G)} \left(\bar{d}(x)\bar{d}(y)\right)^b \geq w(G) \log^{-4b} t, \end{aligned}$$

where G is any graph with vertices $1, 2, \ldots, t$ and the degrees satisfying the bounds (12).

If we view G^* as obtained by swapping edges in $G = G(t)$, then it is easy to see that each such swap increases $\bar{w}(G)$. Indeed, if $u > x > y > v$, then

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

implying that

$$\begin{aligned} &(\bar{d}(x))^b (\bar{d}(v))^b + (\bar{d}(u))^b (\bar{d}(y))^b \\ &> (\bar{d}(x))^b (\bar{d}(y))^b + (\bar{d}(u))^b (\bar{d}(v))^b. \end{aligned}$$

Therefore, $\bar{w}(G^*) \geq \bar{w}(G(t))$.

Next we derive an upper bound on $\bar{w}(G^*)$. Because of the greedy process of adding edges to G^*, a vertex x in G^* has "incoming" edges which originate from vertices $\mathrm{first}(x), \mathrm{first}(x)+1, \ldots, \mathrm{last}(x)$. All m edges outgoing from each vertex $y = \mathrm{first}(x)+1, \ldots, \mathrm{last}(x)-1$ point to x. Thus

we have

$$\begin{aligned}\bar{w}(G^*) &= 2\sum_{\{y,x\}\in E(G^*)}\left(\bar{d}(x)\bar{d}(y)\right)^b \\ &\le 2\sum_{x=1}^{t}\sum_{y=\text{first}(x)}^{\text{last}(x)} m\left(\bar{d}(x)\bar{d}(y)\right)^b \\ &\le 2\sum_{x=1}^{t} d(x)\left(\bar{d}(x)\bar{d}(\text{first}(x))\right)^b \\ &\le 2\log^2 t\sum_{x=1}^{t}\left(\bar{d}(x)\right)^{1+b}\left(\bar{d}(\text{first}(x))\right)^b. \quad (13)\end{aligned}$$

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

$$\begin{aligned}2m\cdot\text{first}(x) &\ge m\cdot\text{first}(x)+mx \ge \sum_{z=1}^{x-1} d(z) \\ &\ge \sum_{z=1}^{x-1}\left(\frac{t}{z}\right)^{(1-\epsilon)/2} \ge t^{(1-\epsilon)/2}x^{1-(1-\epsilon)/2}.\end{aligned}$$

Thus

$$\begin{aligned}\bar{d}(\text{first}(x)) &= \left(\frac{t}{\text{first}(x)}\right)^{1/2} \\ &\le \left(\frac{2mt}{t^{(1-\epsilon)/2}x^{1-(1-\epsilon)/2}}\right)^{1/2} \\ &= (2m)^{1/2}\left(\frac{t}{x}\right)^{(1+\epsilon)/4}. \quad (14)\end{aligned}$$

Using (13) and (14), we get

$$\begin{aligned}\bar{w}(G^*) &\le 2(2m)^{b/2}\log^2 t\sum_{x=1}^{t}\left(\frac{t}{x}\right)^{(1+b)/2}\left(\frac{t}{x}\right)^{b(1+\epsilon)/4} \\ &= 2(2m)^{b/2}\log^2 t\sum_{x=1}^{t}\left(\frac{t}{x}\right)^{1/2+(3/4)b(1+\epsilon)}. \quad (15)\end{aligned}$$

Choosing b so that

$$1/2+(3/4)b(1+\epsilon)\le 1, \quad (16)$$

the sum in (15) is $O(t\log t)$, and we have

$$w(G)\le\log^{4b}\bar{w}(G)\le\log^{4b}\bar{w}(G^*)=O(t\log^{4b+3}t).$$

Proceeding now as in the proof of Theorem 2, we get the similar bound on the seeded cover time of $S(a)$:

$$C^*_{S(a)} = O\left(t^{1-2ba(1-\epsilon)}\,\text{polylog}\,t\right).$$

We take $b=\frac{2}{3}(1-\epsilon)$ to satisfy (16) and obtain

$$C^*_{S(a)} = O\left(t^{1-(4/3)a(1-\epsilon)^2}\right).$$

Similarly as in the proof of Theorem 2, we can conclude that all vertices in $S(a)$ are discovered in $O(t^{1-(4/3)a(1-\delta)})$ steps **whp**, for $\delta=3\epsilon$, and that the cover time of the graph $G(t)$ is $O(t\,\text{polylog}\,t)$.

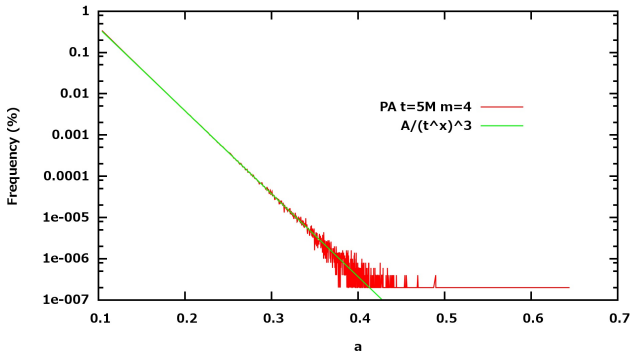

Figure 1: Degree Distribution Of $G(t)$

8. EXPERIMENTAL RESULTS

8.1 Preferential Attachment Graph

Theorem 1 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 $G(m,t)$. 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 $t=5.10^6$ vertices. The representative degree distribution of such 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. $a=\log d/\log t$, while the y-axis is the frequency of the vertices of degree t^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 $b=1/2$. 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 $-a$, $-3a/2$ and $-2a$, are included for discussion purposes only, and the intercepts have no meaning. It is worth noting that the cover times plotted are the average cover times of 10 runs of each of the methods.

Before discussing Figure 2 in greater detail, we remark that it broadly confirms the implications from our theoretical analysis: for random preferential attachment graphs, biassed random walks discover quickly all higher degree vertices while not increasing by much the cover time of the whole graph. For example, by checking the exact cover times, we observed that the biassed random walk with $b=1/2$ took on average 2.7 times longer than a simple random walk to cover the whole graph $G(4,5.10^6)$, but discovered the 100 highest degree vertices 10 times faster than a simple random walk.

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 [13]. The intercept of the y-axis at $m=4$ predicted by this is $1+(\log(\frac{8}{3}\log 5.10^6))/(\log 5.10^6)=1.24$ and this agrees well with the experimental intercept of 1.23. This agreement helps confirm our experimental results.

For a weighted random walk, the stationary distribution

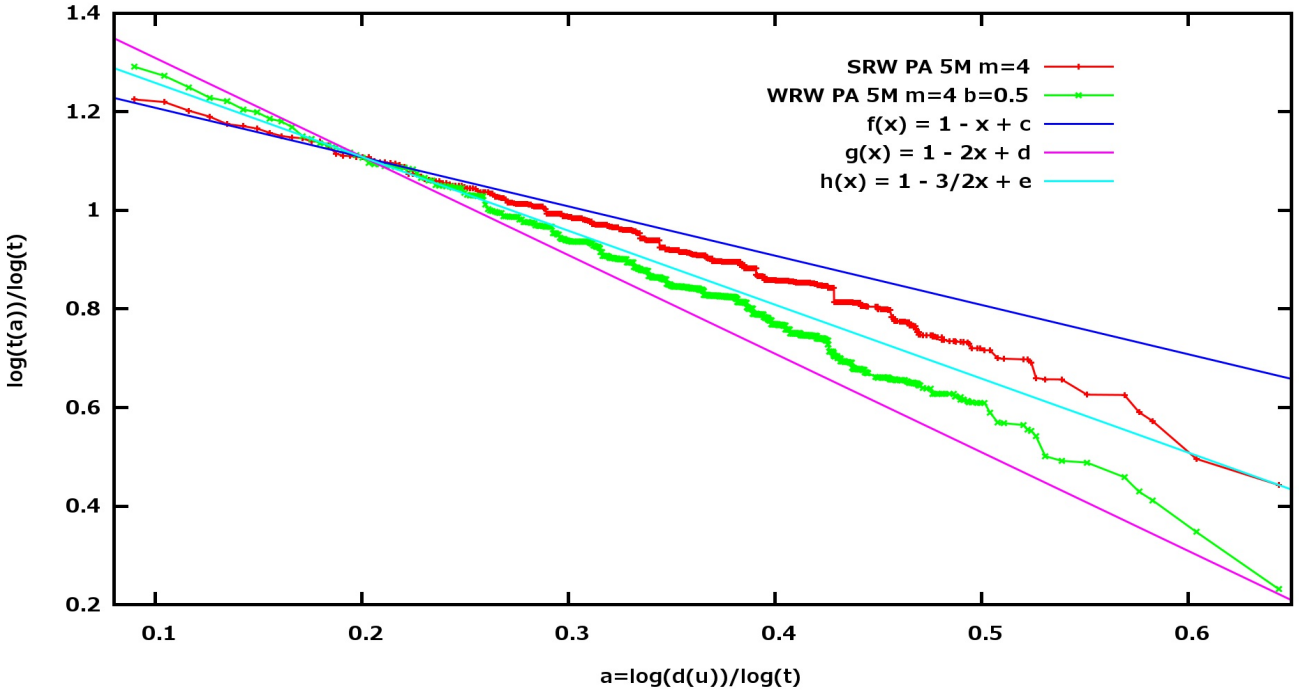

Figure 2: Plots of experimental data showing cover time of all vertices of degree at least t^a as a function of a

$\pi(v)$ of vertex v is given by

$$\pi(v) = \frac{1}{w(G)} \sum_{x \in N(v)} w(v,x),$$

where $w(G)$ is the sum of the edge weights of G, each edge counted twice. Thus for a simple random walk on $G(m,t)$, $\pi_S(v) = d(v)/2mt$. For the weighted random walk of Theorem 1 ($\eta = 1/2$ and $b = 1/2$ for $G(m,t)$) we have the following lower bound.

$$\pi_W(v) = \Omega((d(v))^{3/2}/t\log^5 t).$$

This bound holds because we know from (10) that $w(G) = O(t\log^5 t)$, and

$$\begin{aligned}\sum_{x\in N(v)} w(v,x) &= \sum_{x\in N(v)} (d(v)d(x))^{1/2} \\ &\geq (d(v))^{1/2} \sum_{x\in N(v)} (m)^{1/2} \\ &\geq (d(v))^{3/2}.\end{aligned}$$

We can give an informal explanation of Figure 2 as follows. In the long run, the number of visits to vertex v in T steps approaches $T\pi(v)$, so the first visit to v should be at about $T(v) = 1/\pi(v)$. As $\pi(v)$ increases with increasing degree $d(v)$, then if $h > a$ we should expect to see all vertices of degree t^h before all vertices of degree t^a.

For a simple random walk, let v be a vertex of degree t^a, then $T(v) \approx 1/\pi_S(v) = 2mt/t^a \approx t^{1-a}$. So the SRW plot in Figure 2 should have slope $-a$, and this is indeed the case.

For a weighted random walk, the same argument gives

$$\frac{1}{\pi_W(v)} = O\left(\frac{t\log^5 t}{(t^a)^{3/2}}\right) = \tilde{O}(t^{1-3/2a}),$$

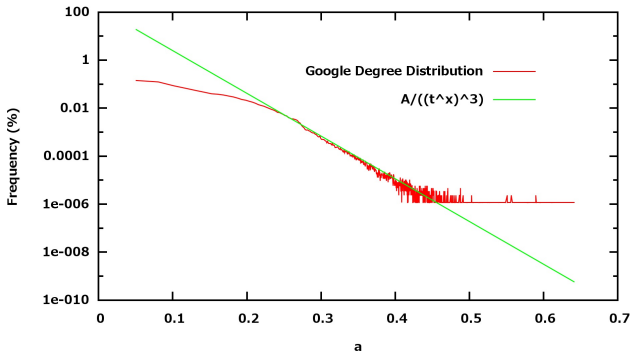

Figure 3: Degree Distribution Of the Google Web-Graph

which explains the slope of $-3a/2$ for the WRW plot.

The total number $n(a)$ of vertices of degree at least t^a is approximated by $\sigma = t^{1-a/\eta} = t^{1-2a}$, where the value of σ from (9), is the expected step at which a vertex of degree t^a is added, and $\eta = 1/2$ for preferential attachment. As no walk based process can visit σ vertices in less than σ steps, this explains the line with slope $-2a$ in Figure 2.

8.2 Real World Networks

In this section we present our experimental results on a real world graphs: a sample of the Google web-graph and the SlashDot Zoo (November 2008 dataset). The dataset for these networks was obtained through the Stanford Network Analysis Project site which contains a wide range of

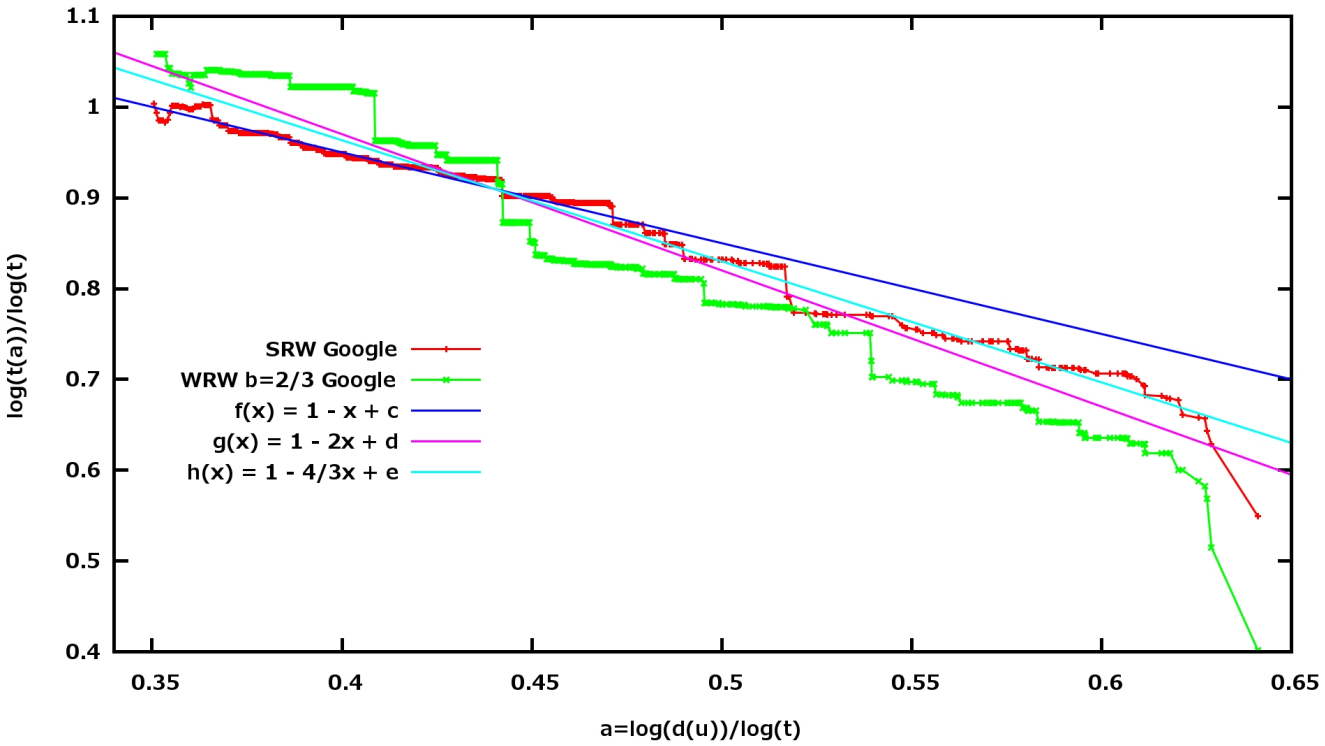

Figure 4: Plots of experimental data showing cover time of all vertices of degree at least t^a (ignoring $a < 0.35$) as a function of a in a sample of the Google web-graph

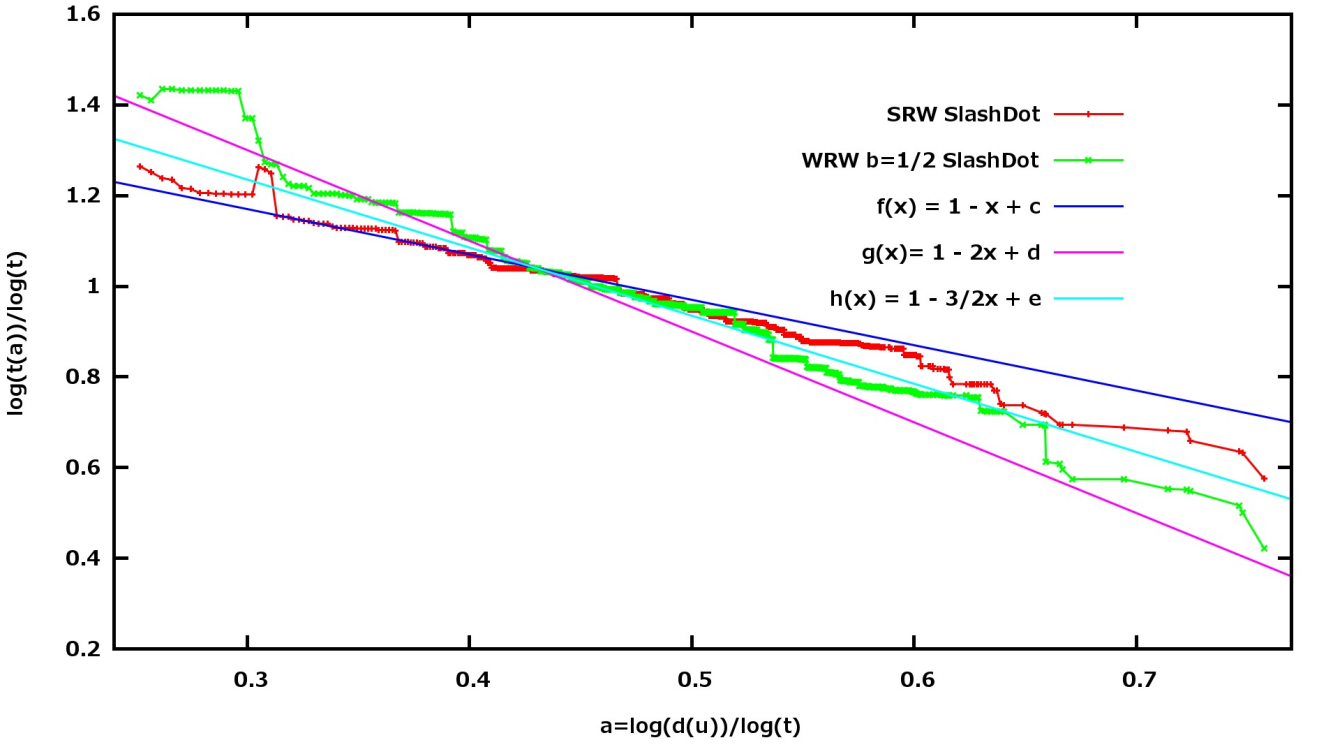

Figure 5: Plots of experimental data showing cover time of all vertices of degree at least t^a (ignoring $a < 0.25$) as a function of a in the SlashDot graph.

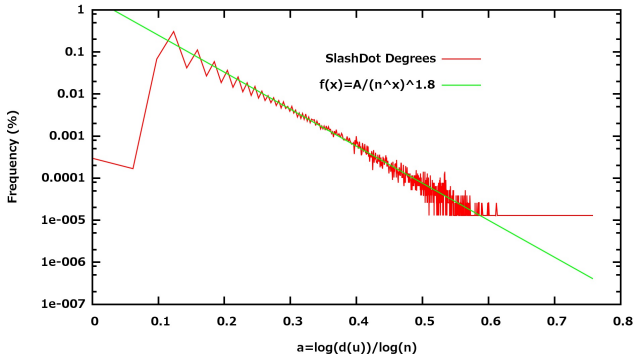

Figure 6: Degree Distribution Of the SlashDot

resources as well as data sets.[1] While both these graphs are directed we are ignoring edge direction and treating them as undirected for the purpose of our experiments. In the case of the Google dataset we only take into account the largest weakly connected component while SlashDot is already weakly connected.

The google web-graph sample has a power-law degree distribution in the mid-range with a co-efficient close to $c = 3$, as seen in Figure 3. As we can see in Figure 4 our method outperforms a SRW in discovering all high degree vertices giving us a strong indication of the effectiveness of our method even on real world networks. The value of b used in this case was $b = \frac{2}{3}$.

In addition as we can see in Figure 6 the degree distribution of the SlashDot graph follows a power-law, with a co-efficient of approximately 1.8. This is lower than the power-law range in which our method was proven to work. However as it is seen in Figure 5 the biassed random walk is still quicker to cover all high degree vertices than a SRW. The value of b used for this case was $b = \frac{1}{2}$. In both cases what is plotted is an average of the cover times of 10 runs of each method.

9. CONCLUSIONS

We have analysed the number of steps required by biassed random walks to discover all higher degree vertices in random t-vertex preferential attachment graphs, and we have proven sublinear upper bounds for discovering all vertices with degree at least t^a, for $0 < a < 1/2$. Our experimental results confirm the good performance of biassed random walks on such graphs. Our theoretical analysis applies also to generalized web-graph processes.

Our theoretical bounds are probably not tight and it would be interesting to see if better bounds can be proven. What is the best value for the parameter b of biassed random walks? From the practical point of view, it would be interesting the investigate the performance of biassed random walks on additional real networks which exhibit the power law.

10. 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 *Proceedings of the 14th international conference on World Wide Web / Industrial and Practical Experience Track*, 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 and O. Riordan. *Handbook of Graphs and Networks: Mathematical results on scale-free graphs*, pages 1–32. S. Bornholdt, H. Schuster (eds), Wiley-VCH, 2002.

[6] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24:5–34, 2004.

[7] 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.

[8] M. Brautbar and M. Kearns. Local algorithms for finding interesting individuals in large networks. In *Proceedings of ICS 2010*, pages 188–199, 2010.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *Proceedings of the Ninth International World-Wide Web Conference WWW9*, Amsterdam 2000. http://www.www9.org/w9cdrom/160/160.html.

[10] C. Cooper. Classifying special interest groups in web graphs. In *Proc. RANDOM 2002: Randomization and Approximation Techniques in Computer Science*, pages 263–275, 2002.

[11] C. Cooper. The age specific degree distribution of web-graphs. *Combinatorics Probability and Computing*, 15:637–661, 2006.

[12] C. Cooper and A. Frieze. A general model web graphs. In *Random Structures and Algorithms*, vol. 22(3), pages 311–335, 2003.

[13] C. Cooper and A. Frieze. The cover time of the preferential attachment graphs. *Journal of Combinatorial Theory*, B(97):269–290, 2007.

[14] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Internet Technol.*, 2:205–223, August 2002.

[15] M. E. E. Drinea and M. Mitzenmacher. Variations on random graph models for the web. Tech. report, Harvard University, Dept. of Computer Science, 2001.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '99, pages 251–262, New York, NY, USA, 1999. ACM.

[1] http://snap.stanford.edu/data/index.html datasets retrieved 15/12/2011. SlashDot consists of 77360 vertices and 905468 edges while the largest WCC of the google sample consists of 855802 vertices and 5066842 edges

[17] 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.

[18] 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.

[19] 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.

[20] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: measurements, models, and methods. In *Proceedings of the 5th Annual International Conference on Computing and Combinatorics*, COCOON'99, pages 1–17, 1999.

[21] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2006.

[22] L. Lovász. Random walks on graphs: A survey. *Bolyai Society Mathematical Studies*, 2:353–397, 1996.

[23] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. In *Reviews Of Modern Physics*, vol. 74, pages 47–97, 2002.

[24] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Evaluating sampling techniques for large dynamic graphs. Technical Report CIS-TR-08-01, Department of Computer and Information Science, University of Oregon, September 2008.

[25] S. Redner. How popular is your paper? An empirical study of the citation distribution. In *European Physical Journal B* vol. 4(2), pages 131–134, 1998.

[26] 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.