A Random-Surfer Web-Graph Model

Avrim Blum* T-H. Hubert Chan* Mugizi Robert Rwebangira*

Abstract

In this paper we provide theoretical and experimental results on a random-surfer model for construction of a random graph. In this model, a new node connects to the existing graph by choosing a start node at random and then performing a short random walk. We show that in certain formulations, this results in the same distribution as the preferential-attachment random-graph model, and in others we give a direct analysis of power-law distribution of degrees or "virtual degrees" of the resulting graphs. We also present experimental results for a number of settings of parameters that we are not able to analyze mathematically.

1 Introduction

There has been substantial work in recent years on the preferential attachment random graph model. In this model, a graph is constructed in the following manner. Nodes arrive one at a time, and each new node makes k connections to the existing graph. However, unlike classic random graph models, these connections are not made uniformly at random, but rather with probability proportional to the degree of existing nodes in the graph. This process is known to produce graphs with a power law degree distribution [2] and that have high conductance [15], and has been proposed as a model for graphs such as the graph of links between pages on the World Wide Web.

A natural question that arises when considering the preferential attachment model is *why*: why should a new node connect to existing nodes with probability proportional to their degree? Is it because we imagine that high degree nodes are "better" (and the degree of a node is an indicator of its quality) or is it for some other reason?

The starting point for this paper is the observation that a simple "random surfer" model provides a natural explanation for preferential attachment. In particular, imagine that each new node (a person setting up their web page) puts in k links into the existing graph by picking a random start node and then randomly surfing the web until it finds k interesting pages to connect to. Imagine also that each page is equally likely to be interesting to the surfer and each link is bidirectional (so we have an undirected graph). Then, if the probability p of a page being "interesting" is sufficiently small, these connections will be made (approximately) according to the stationary distribution of the walk, which is exactly the preferential attachment distribution. Furthermore, since such graphs have high conductance [15], one should not need an extremely low value of p for this to hold. Thus, preferential-attachment may arise even if all nodes are in a sense "equally good", and differences between degrees may not necessarily be an indicator of differences in inherent quality.

Based on this as motivation, in this paper we propose and analyze several "random surfer" models for graph construction. We also give a number of experimental results, both for models we know how to analyze and for several that we do not. Interestingly, the models we are best able to analyze in this setting are all *directed* graph models, rather than undirected models as the one described above. In addition, some of these models can be thought of as making a bridge between the preferential attachment model and the copying model of [13].

2 Random Surfer Models

In this section, we describe several random surfer models that we will examine in the rest of the paper. In each model, nodes arrive one at a time, making k connections to the existing graph. In some models these connections will be viewed as directed edges, and in some as undirected edges. All our models begin with a single start node v_0 having k self-loops. In general, we use v_t to denote the vertex added in the t^{th} step, and n as the total number of vertices.

To motivate our first model, note that if the connections to the existing graph are made uniformly at random, then we have an online version of the standard Erdos-Renyi graph model, and with high probability the maximum degree will be $O(\log n)$. On the other hand, suppose we make each connection by first picking a random start node in the existing graph, and then taking a random walk of *exactly one step*. Then, in the directed case, this will just produce a star (all edges will point to the root v_0), and in the undirected case, it is not hard to show that there is a good chance this produces something star-like of maximum degree $\Omega(n)$.[1]

*Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213. {*avrim, hubert, rweba*} *at cs.cmu.edu.*

[1] In particular, if the graph is currently a star of t nodes, then there is a $(t-1)/t$ chance the random start node is one of the spokes, so the 1-step walk moves to the center and the next edge maintains the star. More generally, with high probability, the number of non-leaf vertices remains small and the expected degree of the initial node is $\Omega(n)$ See Section 3.3.

However, if we flip a coin and with probability $p \in (0,1)$ connect to the random start and with probability $1-p$ take a 1-step walk, then we get something much more natural.

MODEL 1. (1-STEP WALK WITH SELF-LOOP) *In this model, we are given parameters k and p. At time t, vertex v_t makes k connections to the existing graph by repeating the following process k times:*

1. *Pick an existing node v uniformly at random from $\{v_0, \ldots, v_{t-1}\}$.*
2. *With probability p stay at v; with probability $1-p$ take a 1-step walk to a random neighbor of v.*
3. *Add an edge from v_t to the current node.*

In the directed version, the edges added are directed from v_t into the existing graph. In the undirected version, edges are undirected.

Our next model is a walk of the form given in the Introduction: instead of taking one step, we keep walking until we find a node of interest and then connect there. In order to make the model easier to think about, for the case $k > 1$ we imagine after each connection we re-start at a new random start node when performing the next walk.

MODEL 2. (RANDOM WALK WITH COIN FLIPS) *In this model, we are again given parameters k and p. At time t, vertex v_t makes k connections to the existing graph by repeating the following process k times:*

1. *Pick an existing node v uniformly at random from $\{v_0, \ldots, v_{t-1}\}$.*
2. *Flip a coin of bias p*
3. *If the coin comes up heads add an edge from v_t to the current node and stop.*
4. *If the coin comes up tails, move to a random neighbor of the current node and go back to (2).*

In the directed version, the edges added are directed from v_t into the existing graph. In the undirected version, edges are undirected.

3 Theoretical results

3.1 Directed Walk with Self-Loop. Our first (simple) result is that the directed version of Model 1 with $p = 1/2$ is exactly the preferential attachment model.

THEOREM 3.1. *The directed version of Model 1, with $p = 1/2$, has the same distribution as preferential attachment.*

Proof. First, notice that the graph is necessarily a DAG, with all edges pointing backwards in time, and each vertex has an out-degree of k. Now, consider some vertex u in the existing graph with in-degree d_u. An edge from the new vertex v_t will connect to u if either the process chooses u as the start node of its walk and does not take a step, or else it chooses one of u's in-neighbors u' as the start node and *does* take a step, selecting the edge from u' to u. The first case has probability p/t, and the second case has probability $(1-p)d_u/(kt)$. For $p = 1/2$, the sum of these two quantities is $(k + d_u)/(2kt)$ which is exactly proportional to the total degree $k + d_u$ of u. ■

One implication of Theorem 3.1 is that for $p > 1/2$, the model is a mixture of preferential-attachment and uniform-random connections. That is, the case $p > 1/2$ can be viewed as: with probability $2p-1$ choose a neighbor uniformly at random, and with the remaining probability choose a neighbor with probability proportional to degree. This process is known to produce power-law degree distributions. For general $p \in (0,1)$, we now give an argument for power-law degree distributions from first principles.

Let $d_i(t)$ be the number of nodes with in-degree i at step t, and $D_i(t)$ be the expectation of $d_i(t)$. We now analyze $D_i(t)$ via the following equation.

$$D_i(t+1) = D_i(t) + \tag{3.1}$$

$$\frac{pk}{t} \cdot \{D_{i-1}(t) - D_i(t)\} + \tag{3.2}$$

$$\frac{(1-p)k}{t} \cdot \{(i-1)D_{i-1}(t) - iD_i(t)\} \cdot \frac{1}{k}. \tag{3.3}$$

Observe that the number of nodes with in-degree i increases if the new node connects to an existing node of degree $i-1$ and decreases if the new node connects to one of degree i. The term in (3.2) is due to the fact that with probability p the new node is connected to an existing node picked uniformly at random. The term in (3.3) corresponds to the case when with probability $1-p$, the new node connects to a random out-going neighbor of a randomly picked node. The factor k appears in both (3.2) and (3.3) because each new node makes k connections to the existing nodes. The factor $1/k$ appears only in (3.3) because in the case where a random out-going neighbor is chosen, there are k possible choices. We require for large enough t, a new node does not make more than one connection to an existing node.

THEOREM 3.2. *There exists a constant $C > 0$ such that as t tends to infinity, $D_i(t) \sim Ci^{-\frac{2-p}{1-p}}t$.*

Proof. Using the above equations, the proof follows directly from the techniques of Kumar et al. [13], Cooper and Frieze

[10], and Mitzenmacher [16], which allow one to determine the asymptotic behavior of $D_i(t)$.

In particular, for each i, we make the substitution $D_i(t) = c_i t$ in (3.1) - (3.3) to obtain the following equation.

$$c_i = pk \cdot \{c_{i-1} - c_i\} + (1-p) \cdot \{(i-1)c_{i-1} - ic_i\} \quad (3.4)$$

Rearranging (3.4), we have

$$\frac{c_i}{c_{i-1}} = 1 - \frac{2-p}{1+pk+(1-p)i} \sim 1 - \frac{2-p}{1-p} \cdot \frac{1}{i},$$

for large values of i. Using the fact that $\Pi_{i=1}^{n}(1+\lambda/i) = \Theta(n^{\lambda})$, we have

$$c_i = \Theta(\Pi_{j=1}^{i}(1 - \frac{2-p}{1-p} \cdot \frac{1}{j})) \sim Ci^{-\frac{2-p}{1-p}},$$

for some $C > 0$. ■

Moreover, using Theorem 4 of [10], one can also show that $d_i(t)$ is concentrated around its mean, as stated in the following theorem.

THEOREM 3.3. *For any $\rho > 0$,*

$$Pr(|d_i(t) - D_i(t)| \geq \rho) \leq \exp(-\frac{\rho^2}{8kt}).$$

3.2 Directed Walk with Coin Flipping. We now consider the directed case of Model 2, for the case $k = 1$. That is, we connect a new node to the existing graph by picking a start node u uniformly at random, and then performing a random walk, where at each step we halt the walk with probability p. Since $k = 1$, we can view the random graph constructed as a tree, in which the initial node is the root and every other node has an edge directed to its parent.

To analyze this walk, we define a notion of the *virtual degree* of a node that is related to the node's actual degree, but also contains terms for the local neighborhood of the node as well. We then prove that for this definition, at each step the expected increase in virtual degree of any given node is proportional to the virtual degree itself. (The virtual degree itself is a fractional quantity, and at each step will change by at most some constant.) Using this, we can show that the expected virtual degrees follow a power-law, and we can also give some bounds on their concentration about their means. Moreover, we can give a crude lower bound on the expected *real* degree of a given node, which is comparable to its expected virtual degree.

However, our concentration bounds are not sharp enough to give a true proof that the virtual degrees, or the real degrees, follow the power law.

DEFINITION 1. *Suppose u is a node in the tree. For $i \geq 0$, denote $L_i(u)$ to be the set of level i descendants of u and $l_i(u) = |L_i(u)|$. For instance, $L_0(u)$ is the set of children, $L_1(u)$ is the set of grandchildren, and so on. Let $\beta = \{\beta_i\}_{i \geq 0}$ be a sequence of real numbers such that $\beta_0 = 1$. The virtual degree of u with respect to β is*

$$\nu(u) = 1 + \sum_{k \geq 0} \beta_k l_k(u).$$

In the definition of virtual degree $\nu(u)$, the leading term 1 corresponds to the parent of u. We require $\beta_0 = 1$, for each child of u should contribute 1 towards the degree of v. We would like the virtual degree to reflect the actual degree of a node, and hence ideally, for $i \geq 1$, we would like β_i to be small. On the other hand, we also want that the expected increase in the virtual degree $\nu(u)$ of node u in each step to be proportional to its current virtual degree. The following theorem states we can satisfy these requirements simultaneously.

THEOREM 3.4. *Suppose we consider the directed walk with coin flipping probability $p \in (0,1)$. Then, there exists $\beta = \{\beta_k\}_{k \geq 0}$, dependent on p, with $\beta_0 = 1$ such that for each node u, the expected increase in $\nu(u)$ from step t to step $t+1$ is $p/t \cdot \nu(u)$. Moreover, for $k \geq 0$, $|\beta_k| \leq 1$, and as k tends to infinity, β_k tends to zero exponentially, i.e. there is some $C > 0$ and $0 < \rho < 1$ such that $|\beta_k| \leq C\rho^k$.*

Proof. We fix the coin flipping probability p and find some sequence β that satisfies the requirements.

For convenience, we denote $q = 1 - p$ and $L_{-1}(u) = \{u\}$. Then, for $i \geq 0$, if a new connection is made to a node in $L_{i-1}(u)$, then the increase in $\nu(u)$ is β_i.

Fix $i \geq 0$. We first calculate the probability that a new connection is made to a node in $L_{i-1}(u)$. Recall that we first pick a node uniformly at random to start the directed random walk. If we end up making a new connection to a node in $L_{i-1}(u)$, we must have begun the random walk at some node in $L_{i-1+j}(u)$, for some $j \geq 0$.

We fix some $j \geq 0$ and calculate the probability that the random walk starts at some node in $L_{i-1+j}(u)$ and ends up at some node in $L_{i-1}(u)$. Note that there are $l_{i-1+j}(u)$ nodes to start and there are j hops to be made. Hence, the probability is $l_{i-1+j}(u)/t \cdot q^j \cdot p$.

It follow that the probability that a new connection is made to some node in $L_{i-1}(u)$ is $\frac{p}{t}\sum_{j \geq 0} q^j l_{i-1+j}(u)$.

Hence, the expected increase in $\nu(u)$ from step t to step $t+1$ is

$$\begin{aligned}
&\sum_{i\geq 0} \beta_i \cdot \frac{p}{t} \sum_{j\geq 0} q^j \, l_{i-1+j}(u) \\
= \;& \frac{p}{t} \sum_{i\geq 0} \sum_{k\geq i-1} \beta_i \, q^{k-i+1} \, l_k(u) \\
= \;& \frac{p}{t} \sum_{k\geq -1} \sum_{0\leq i\leq k+1} \beta_i \, q^{k-i+1} \, l_k(u)
\end{aligned}$$

Recall we wish that the above quantity to be equal to

$$\frac{p}{t}\nu(u) = \frac{p}{t} \cdot \{1 + \sum_{k\geq 0} \beta_k \, l_k(u)\}.$$

Hence, it suffices to find a sequence β such that the corresponding coefficients of $l_k(u)$ are equal.

For $k = -1$, we require $\beta_0 = 1$; for $k = 0$, we have $\beta_0 q + \beta_1 = \beta_0$, which implies that $\beta_1 = p$. In general, for $k \geq 0$, we have

$$\beta_k = \sum_{0\leq i\leq k+1} \beta_i \, q^{k-i+1}.$$

Now, suppose $k \geq 0$. Then, we have

$$\begin{aligned}
\beta_{k+1} &= \sum_{0\leq i\leq k+1} \beta_i \, q^{k-i+1} \\
&= \beta_{k+2} + q \sum_{0\leq i\leq k+1} \beta_i \, q^{k-i+1} \\
&= \beta_{k+2} + q\beta_k.
\end{aligned}$$

Hence, the sequence β can be determined by the recurrence $\beta_0 = 1$, $\beta_1 = p$ and for $k \geq 0$, $\beta_{k+2} - \beta_{k+1} + q\beta_k = 0$.

We show inductively that $|\beta_k| \leq 1$. We first observe that this is true for $k = 0, 1, 2$. Assume that the result is true for integers up to $k+1$. In the first case, suppose β_k and β_{k+1} have the same sign. Then, $|\beta_{k+2}| = ||\beta_{k+1}| - q|\beta_k|| \leq 1$, by the induction hypothesis. In the second case, suppose β_k and β_{k+1} have different signs. Hence, $|\beta_{k+2}| = |\beta_{k+1} - q\beta_k| \leq |\beta_{k+1} - \beta_k| = q|\beta_{k-1}| \leq 1$, by the induction hypothesis.

For $p = 3/4$, we have $\beta_k = \frac{k+2}{2^{k+1}}$. Otherwise, for other values of p in $(0, 1)$, let $\lambda_1 = (1 - \sqrt{1-4q})/2$ and $\lambda_2 = (1 + \sqrt{1-4q})/2$ and $\beta_k = A\lambda_1^k + B\lambda_2^k$, for some constants A and B. Observe that since $0 < p < 1$, the magnitudes of λ_1 and λ_2 are both strictly less than 1. Hence, in any case, as k tends to infinity, β_k tends to 0 exponentially. ■

For the rest of the discussion, we consider the virtual degree defined with respect to some sequence β that satisfies Theorem 3.4. We next explore how the virtual degree of a particular node changes with time. Define $\nu_t(u)$ to be the virtual degree of node u at step t and t_u to be the time when node u first appears. Then, it follows that $\nu_{t_u}(u) = 1$, since each new node is a leaf when it first appears.

THEOREM 3.5. *For any node u and step $t \geq t_u$, the expectation $E[\nu_t(u)] = \Theta((t/t_u)^p)$.*

Proof. For any $t > t_u$, we have from Theorem 3.4 that

$$E[\nu_t(u)] = (1 + p/(t-1)) \, E[v_{t-1}(u)].$$

Hence,

$$E[\nu_t(u)] = \Pi_{i=t_u}^{t-1}(1 + p/i) = \Theta((t/t_u)^p).$$

■

We next give an intuition, similar in spirit to [3], of how Theorem 3.5 suggests that the virtual degrees of the random graph should follow the power law. Suppose the random process is run for n steps to form a random graph with n nodes. Then, from Theorem 3.5, the expected virtual degree of the ith node joining the graph is $\Theta((n/i)^p)$. If we let $\kappa \approx \Theta((n/i)^p)$, we would have $i \approx \Theta(n\kappa^{-1/p})$. Observing that nodes joining later should probably have smaller virtual degrees, one might expect that the proportion of nodes having virtual degrees smaller than κ to be $1 - \Theta(\kappa^{-1/p})$. Differentiating this quantity with respect to κ, we conjecture that the proportion of nodes having degree κ should be $\kappa^{-(1/p+1)}$.

Unfortunately, we do not have a strong enough concentration bound that would allow us to make the above intuition rigorous. However, using martingale techniques, we can show that the virtual degree cannot be *too* much larger than its mean for the case when the coin flipping probability $p > 1/2$.

THEOREM 3.6. *There exists a constant $C > 0$ such that for coin flipping probability $p > 1/2$ and any $\rho \geq 1$,*

$$Pr[\nu_t(u) \geq C\rho E[\nu_t(u)]] \leq \exp\{-\rho^2/t_u\}.$$

Proof. Consider a node u and recall that t_u is the time when it first appears. Define $a_i = 1 + p/i$. Recall from the proof of Theorem 3.5 that $E[\nu_t(u)] = \Pi_{i=t_u}^{t-1} a_i = \Theta((t/t_u)^p)$.

Define $Y_i = \nu_i(u)/E[\nu_i(u)]$, for $i \geq t_u$. Then, it follows that $\{Y_i\}$ is a martingale. Define $D_i := Y_i - Y_{i-1}$.

Recall that the sequence $\{\beta_k\}$ tends to zero. Hence, it follows that $|\nu_i(u) - \nu_{i-1}(u)| = \Theta(1)$, and we have $|D_i| = |Y_i - Y_{i-1}| = 1/E[\nu_i(u)] \cdot |\nu_i(u) - a_{i-1}\nu_{i-1}(u)| = 1/E[\nu_i(u)] \cdot |\Theta(1) - \frac{p}{i-1} \cdot \nu_{i-1}(u)| = \Theta(1/E[\nu_i(u)])$, since $\nu_{i-1}(u) = O(i-1)$. Hence, we can let $K_i = \Theta(1/E[\nu_i(u)])$, and so $|D_i| \leq K_i$. By the Azuma-Hoeffding martingale inequality, we have for any $x > 0$,

$$Pr[Y_t - Y_{t_u} \geq x] \leq \exp\{-x^2/2 \sum_{i=t_u+1}^{t} K_i^2\}.$$

Observe that for $p > 1/2$, we have

$$\begin{aligned}\sum_{i=t_u+1}^{t} K_i^2 &\leq \sum_{i=t_u+1}^{t} \Theta(1/E[\nu_i(u)]^2) \\ &= \sum_{i=t_u+1}^{t} \Theta((i/t_u)^{-2p}) \\ &= \Theta(t_u(2p-1)\cdot(1-(t/t_u)^{-(2p-1)})) \\ &= \Theta(t_u).\end{aligned}$$

Hence, for some large enough $C' > 0$, if we put $x = C'\sqrt{st_u}$, we have $Pr[Y_t - Y_{t_u} \geq x] \leq e^{-s}$. Observing that $Y_{t_u} = 1$ and taking $\rho = \sqrt{st_u}$, we have

$$Pr[\nu_t(u) \geq C\rho E[\nu_t(u)]] \leq \exp\{-\rho^2/t_u\},$$

where $C > 0$ is a constant large enough to absorb the 1. ■

3.2.1 A Crude lower bound for the expected *real* degree.

Recall that for a given node u in the tree and $i \geq 0$, $L_i(u)$ is the set of level i descendants of u and $l_i(u) = |L_i(u)|$. In particular, $l_0(u)$ is the number of children node u has. We can give a crude lower bound for $l_0(u)$ for any given node u.

THEOREM 3.7. *For any node u and step $t \geq t_u$, the expectation $E[l_0(u)] \geq \Omega((t/t_u)^{p(1-p)})$.*

Proof. Let the number of level i descendants of node u at time step t be $l_i^t(u)$. It follows that

$$\begin{aligned}E[l_0^{t+1}(u)] &= E[l_0^t(u)] \\ &\quad + \frac{p}{t}\cdot\{1 + \sum_{j\geq 0} E[l_j^t(u)](1-p)^{j+1}\} \\ &\geq E[l_0^t(u)] + \frac{p(1-p)}{t}E[l_0^t(u)]\end{aligned}$$

Suppose that for some constant $A > 0$, for some $t > 0$, and α, we have $E[l_0^t(u)] \geq At^\alpha$. Observing that for $t \geq 1$, $(t+1)^\alpha - t^\alpha \leq \alpha t^{\alpha-1}$, we have

$$\begin{aligned}E[l_0^{t+1}(u)] &\geq A\{t^\alpha + \frac{p(1-p)}{t}\cdot t^\alpha\} \\ &\geq A\{(t+1)^\alpha + (p(1-p)-\alpha)t^{\alpha-1}\} \\ &\geq A(t+1)^\alpha,\end{aligned}$$

if we set $\alpha = p(1-p)$.

Note that for $t = t_u + 1$, $E[l_0^t(u)] = \Theta(1)$. Hence, it follows that $E[l_0^t(u)] \geq \Omega((t/t_u)^{p(1-p)})$. ■

3.3 Undirected Walk without Self-loop.

We now consider the model mentioned when motivating Model 1 in which a new connection is made to a random neighbor of a randomly selected node. We show that there is a node, namely the initial node, that in expectation has degree linear in the size of the random tree produced. Thus, the self-loop in Model 1 is crucial for producing natural graphs.

THEOREM 3.8. *Under the undirected walk without self-loop model, the expected number of leaves connected to the initial node in the random tree produced is $\Omega(n)$, where n is the number of nodes.*

Proof. Let L_n be number of leaves connected to the initial node v_0 at step n and D_n be the degree of the initial node v_0 at time n.

Suppose we are at step n. With probability at least L_n/n, a leaf of v_0 would be picked and after one jump, a new connection would be made to v_0, causing the number of leaves connecting to v_0 to increase by 1. On the other hand, with probability $\frac{1}{n}\cdot\frac{L_n}{D_n}$, the initial node v_0 is picked and after one jump a new connection is made to an existing leaf, causing the number of leaves connected to v_0 to decrease by 1.

Hence, $E[L_{n+1} - L_n] \geq L_n/n - 1/n\cdot L_n/D_n \geq 1/n\cdot E[L_n - 1]$, with the last inequality holding because $L_n \leq D_n$. Hence, if we let $Z_n = L_n - 1$, we have $E[Z_{n+1}] \geq (1+1/n)E[Z_n]$. Observe that $E[Z_3] > 0$.

Hence, $E[Z_n] \geq \Pi_{i=3}^{n-1}(1+1/i)E[Z_3] = \Omega(n)$ and so $E[L_n] \geq \Omega(n)$. ■

4 Experimental results

All experiments were the average of 100 runs with a size $n = 100,000$ nodes and $k = 1$, i.e. the random graph produced is a tree. In each case, we investigate how the average proportion P_d of nodes having degree d varies with d. Since we wish to observe whether the degree distribution follows a power law, we plot $\log_{10} P_d$ against $\log_{10} d$, for d up to 40. All four models exhibits power-law like phenomenon. Figure 5 shows the degree distribution for the four models and they behave similarly, although the maximum degree seen is much larger for the directed models than for the undirected ones.

4.1 Directed walk with self-loops.

Figure 1 shows experimentally that the power-law phenomenon exhibited by the degree distribution becomes more apparent as the probability p decreases and the degree d increases. Notice that for $p = 1$, this is just the Erdos-Renyi random graph model, which does not obey the power law. Moreover, the maximum degree seen for $p = 1$ is only about 20. As p gets smaller the graph can be fitted better with a straight line. On the hand, the portion of the graph corresponding to large degrees can be fitted well with a straight line. Note that even

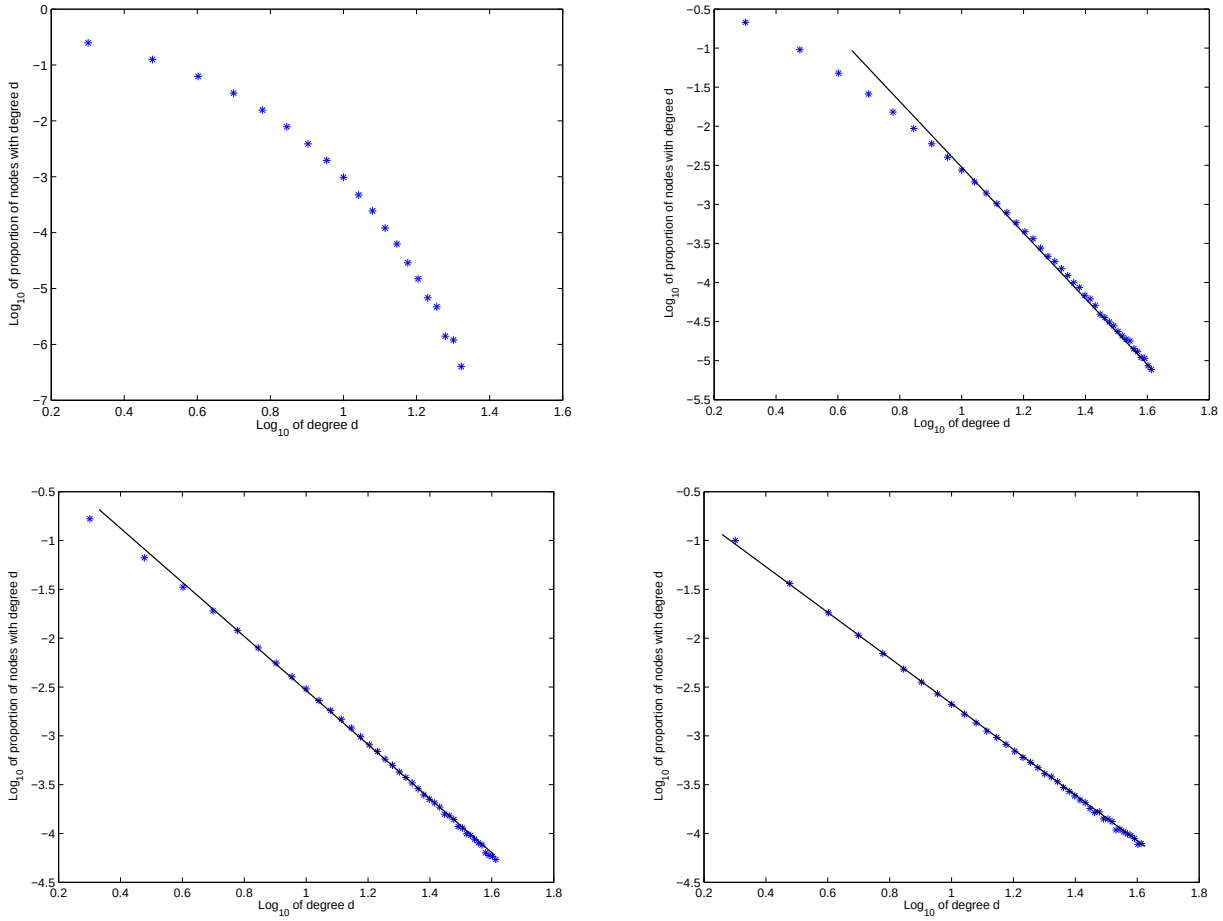


Figure 1: Directed walk with self-loops: (Top-Left) $p = 1$, (Top-Right) $p = 0.75$, (Bottom-Left) $p = 0.5$, (Bottom-Right) $p = 0.25$

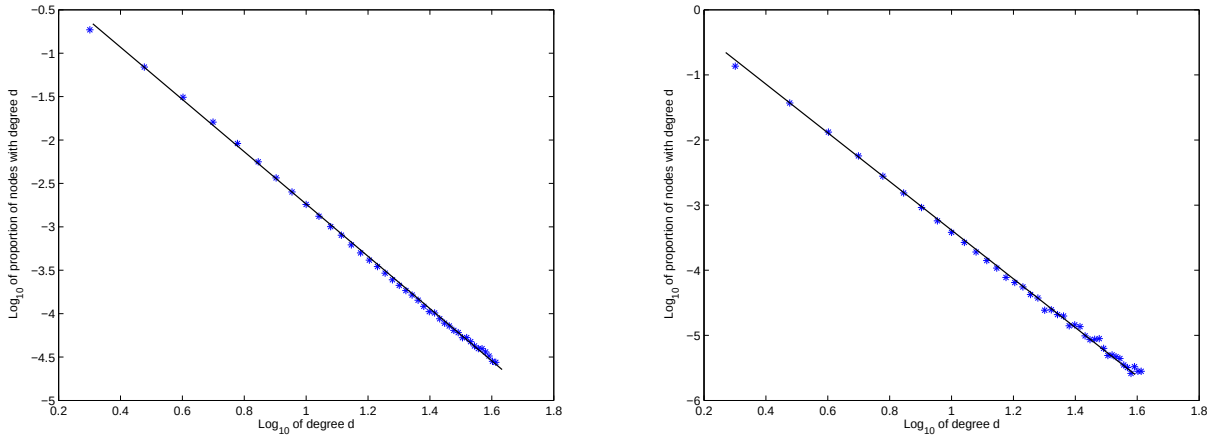


Figure 2: Directed walk with coin flips: (Left) $p = 0.5$ (Right) $p = 0.25$

for $p = 0.75$, power law phenomenon is exhibited for large degrees d.

4.2 Directed walk with coin flips. We do not have a proof, but Figure 2 is very similar to Figure 1, which indicates that in this case the degrees may be following a power law.

4.3 Undirected walk with self-loops. We do not know how to analyze this model yet. As seen in Figure 3, there are indications that power law phenomenon is exhibited by large degrees. On the other hand, the distribution of degrees may follow some other nice distribution that is not very far from power law (e.g. log-normal distribution).

4.4 Undirected walk with coin flips. Like the previous model, this model is not easy to analyze. But Figure 4 shows that the degree sequence does not look too different from undirected walk with self-loops model. We know theoretically that if p is very small the degree sequence will tend closer to a power law. Figure 4 indeed shows that for $p = 0.05$, the graph can be better fitted with a straight line.

5 Conclusions and Open Questions

In this paper we present some initial analysis and experimental results for several simple random-surfer models for web-graph construction. The models are similar in spirit to the copying model of [13], and in fact the directed case of Model 1, for $k = 1$ is identical to both the copying model and preferential-attachment. There are many open questions including:

1. In the case of the directed walk with self-loops, we can analyze the *expected* virtual degrees and provide some concentration bounds, but do not have a formal proof that the virtual degrees necessarily follow a power-law. Furthermore, even assuming this is the case, we do not have a proof that this implies that the actual degrees must be power-law, though our experimental results show this to in fact be the case. Thus, can one give a formal proof that the degrees indeed follow a power law for this model?

2. For the case of the *undirected* walk with self-loops, we know that as p goes to 0, this walk approaches the preferential-attachment distribution. However, experimentally, even for $p = 1/2$ the degrees follow some heavy-tailed distribution. Can one give a formal analysis of the degree distribution in this case?

3. Finally, another issue brought out by this work is that differences between degrees of nodes in the (real) web graph may not necessarily be due to a distinction in quality, but rather just the result of a random walk

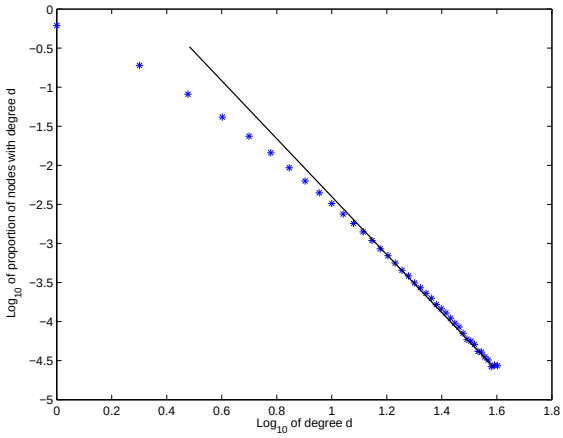
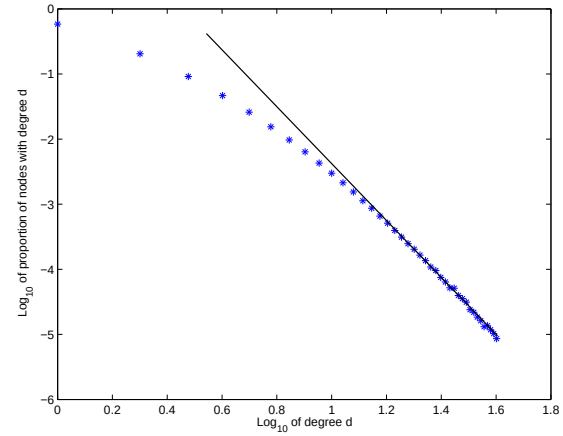


Figure 3: Undirected walk with self-loops: $p = 0.5$



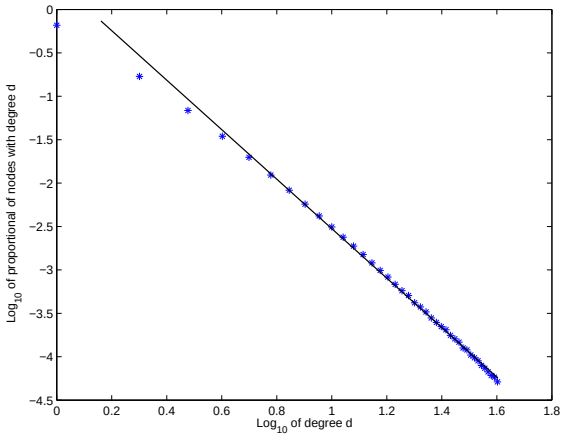


Figure 4: Undirected walk with coin flips: (Top) $p = 0.5$, (Bottom) $p = 0.05$

Degree	Directed walk with self-loops	Directed Walk with coin-flips	Undirected Walk with self-loops	Undirected Walk with coin-flips
1	0.6670	0.6672	0.6136	0.5840
2	0.1669	0.1862	0.1903	0.2044
3	0.06662	0.06929	0.08128	0.09132
4	0.03333	0.03107	0.04137	0.04652
5	0.01900	0.01607	0.02355	0.02596
6	0.01195	0.009108	0.01444	0.01546
7	0.007902	0.005607	0.009301	0.009703
8	0.005547	0.003662	0.006298	0.006354
9	0.004048	0.002524	0.004447	0.004286
10	0.003046	0.001809	0.003242	0.002992
11	0.002332	0.001322	0.002376	0.002134
12	0.001832	0.001006	0.001802	0.001540
13	0.001452	0.0008016	0.001405	0.001131
14	0.001187	0.0006195	0.001088	0.0008657
15	0.0009853	0.0005008	0.0008539	0.0006553
16	0.0007938	0.0004128	0.0006968	0.0005115
17	0.0007005	0.0003486	0.0005608	0.0003950
18	0.0005839	0.0002924	0.0004531	0.0003122
19	0.0005009	0.0002455	0.0003842	0.0002471
20	0.0004400	0.0002118	0.0003121	0.0002031
21	0.0003731	0.0001846	0.0002707	0.0001653
22	0.0003280	0.0001637	0.0002300	0.0001355
23	0.0003001	0.0001426	0.0001990	0.0001082
24	0.0002559	0.0001213	0.0001652	0.0000956
25	0.0002188	0.0001054	0.0001454	0.0000750
26	0.0002020	0.0001018	0.0001289	0.0000639
27	0.0001860	0.0000872	0.0001103	0.0000520
28	0.0001643	0.0000778	0.0000954	0.0000511
29	0.0001545	0.0000720	0.0000851	0.0000395
30	0.0001382	0.0000642	0.0000708	0.0000354
31	0.0001221	0.0000604	0.0000594	0.0000313
32	0.0001116	0.0000528	0.0000564	0.0000240
33	0.0001039	0.0000529	0.0000511	0.0000219
34	0.0000972	0.0000475	0.0000415	0.0000184
35	0.0000904	0.0000425	0.0000407	0.0000162
36	0.0000789	0.0000396	0.0000353	0.0000131
37	0.0000735	0.0000395	0.0000323	0.0000136
38	0.0000649	0.0000362	0.0000264	0.0000118
39	0.0000602	0.0000325	0.0000277	0.0000103
40	0.0000543	0.0000282	0.0000272	0.0000086
Max degree seen in 100 runs	1623	20612	325	138

Figure 5: Average proportion of nodes having different degrees under different models with $n = 100,000$, $p = 0.5$ and 100 runs

process. Thus, if one is using degree as a measure of quality, one may just be picking out nodes that have been around the longest. Instead, some measure that examines the degree of a node *relative* to what one would expect given the time the node has been in the system might be more appropriate.

References

[1] W. Aiello, F.R.K. Chung, and L. LU. A random graph model for massive graphs. *Proc. of the 32nd Annual ACM Symposium on the Theory of Computing*, pages 171–180, 2000.

[2] Reka Albert and Albert-Laszlo Barabasi. Topology of evolving networks: Local events and universality. *Physical Review Letters*, pages 5234–5237, 2000.

[3] Sagy Bar, Mira Gonen, and Avishai Wool. An incremental super-linear internet topology model. *5th annual Passive and Active Measurement Workshop*, 2004.

[4] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, pages 509–512, 1999.

[5] Bollobas and O.Riordan. The diameter of a scale free random network.

[6] Bollobas and O.Riordan. *Handbook of Graphs and Networks*. Wiley VCH, Berln, 2002.

[7] Bollobas, O.Riordan, J.Spencer, and G.Tusanady. The degree sequence of a scale free random graph process. *Random Structures and Algorithms*, pages 279–290, 2001.

[8] F.R.K. Chung, L.LU, and V. Vu. Eigenvalues of random power law graphs. *Annals of Combinatorics*, pages 21–33, 2003.

[9] F.R.K. Chung, L.LU, and V. Vu. The spectra of random graphs with expected degrees. *Proceedings of National Academies of Science*, pages 6313–6318, 2003.

[10] C. Cooper and A. M. Frieze. A general model of undirected web graphs. *Random Structures and Algorithms*, pages 311–335, 2003.

[11] P. Erdos and A. Renyi. On random graphs i. *Publicationes, Mathematicae, Debrecen*, pages 290–297, 1959.

[12] M. Faloutos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*, pages 251–262, 1999.

[13] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. *Proc. IEEE Symposium on Foundations of Computer Science*, 2000.

[14] M. Mihail and C. H. Papadimitriou. On the eigenvalue powerlaw. *Randomization and Approximation Techniques, 6th International Workshop*, pages 254–262, 2002.

[15] M. Mihail, C. H. Papadimitriou, and A. Saberi. On certain connectivity properties of the internet topology. *Proc. IEEE Symposium on Foundations of Computer Science*, 2003.

[16] M. Mitzenmacher. A brief history of generative models for lognormal and power law distributions.

[17] H.A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

[18] Gilbert Strang. *Linear Algebra and its Applications*. Harcourt Brace Jacanovich, 1988.

[19] D.J. Watts. *Small Worlds:They Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, 1999.

[20] G. Yule. A mathematical theory of evolution based on the theories of j.c. willis. *Philosophical Transactions of the Royal Society of London (series B)*, pages 21–87, 1925.