Variations on Random Graph Models for the Web

Eleni Drinea
Mihaela Enachescu
Michael Mitzenmacher

TR-06-01

Computer Science Group
Harvard University
Cambridge, Massachusetts

Variations on Random Graph Models for the Web

Eleni Drinea* Mihaela Enachescu† Michael Mitzenmacher‡

In this paper, we introduce variations on random graph models for Web-like graphs. As a basis, we recall a model first presented in [5]. We add vertices to the graph, one per unit time, with each vertex having one outedge. With probability α this outedge loops back to the new vertex and with probability $1-\alpha$ the end is chosen to be the same as the end of a random extant edge. Notice that in this second case the end of a new edge is chosen proportionally to the current indegrees of the vertices. One feature of these graphs that makes them Web-like is that the indegrees obey a power law; that is, the fraction of vertices of degree i grows like $1/i^{\gamma}$, where in this case $\gamma = 1/(1-\alpha)$.

Model 1: We grow a graph to n vertices, at a rate of one vertex per unit of time. (We may begin initially with t_0 vertices in a directed cycle at time t_0.) Let d_u be the indegree of a vertex u extant at time t. A new vertex v has one directed outedge, with the probability that the end vertex is u is proportional to the weight of u, $w_u = d_u + c$ for a constant $c > 0$. The total weight at time t is $(c+1)t$. One can interpret the constant c as every vertex is given c self-loops, although we do not require that c is an integer. We have found while writing this work that our model is equivalent to the following model from [4]: with probability $c/(c+1)$ the end of a new edge is chosen uniformly at random from extant vertices, and with probability $1/(c+1)$ the end is u with probability proportional to d_u.

We present an argument sketching the asymptotic behavior of the indegrees. (A more rigorous form of this argument based on martingales is given in [4].) Let $n_i(t)$ be the number of vertices extant at time t with indegree i; we write n_i where the meaning is clear. Note

$$E[n_0(t)] = E[n_0(t-1)] + 1 - \frac{cn_0(t-1)}{(c+1)t}.$$

The growth of n_0 is well approximated by the fluid limit

$$\frac{dn_0}{dt} = 1 - \frac{cn_0}{(c+1)t}.$$

Let us assume that in the limit as n gets large that the $n_i(t)$ converge to fixed fractions of the graph, so $n_i(t) = a_i t$. By plugging in the above we find $a_0 = \frac{c+1}{2c+1}$.

More generally, for $i \geq 1$,

$$\frac{dn_i}{dt} = \frac{(i+c-1)n_{i-1} - (i+c)n_i}{(c+1)t}.$$

A simple induction yields $a_i = \frac{i+c-1}{i+2c+1}a_{i-1}$. For integral c, this simplifies to $a_i = (c+1)\binom{2c}{c-1}/((c+2)\binom{2c+i+1}{c+2})$, from which $a_i \sim i^{-(2+c)}$ for large i. Alternatively, we see $a_i/a_{i-1} = 1 - (c+2)/(i+2c+1) \approx ((i-1)/i)^{c+2}$, so $a_i \sim i^{-(2+c)}$ also for non-integral c. If m outedges are produced for each new vertex, a similar argument reveals $a_i \sim i^{-(2+c/m)}$. A similar result can be found using the scale-free analysis of [1]. Hence this simple model allows any power-law exponent greater than 2.

Model 2: The weight of a vertex is $w_u = (d_u + c)^p$ for some constants $c > 0$ and p. The idea behind this model is that the strength of forming a new connection may be proportional to a nonlinear function of the indegree. The limiting cases for this model are interesting: when $p \to \infty$, essentially all edges point to a single node, and when $p \to -\infty$, the graph is essentially a single path. Given recent results on the shape of the Web, showing for example there are many long path-like pieces [3], it is possible that some areas of the Web may be similar to Model 2 with properly chosen parameters.

We note that the differential equation setup used for Model 1 can also be used to gain insight into Model 2, although we lack a closed form solution. Here $\frac{dn_i}{dt} = \frac{(i+c-1)^p n_{i-1} - (i+c)^p n_i}{W(t)}$, where $W(t)$ is the total weight at time t. Assuming also $W(t)$ converges to $W \cdot t$ for some constant W, we have $a_i/a_{i-1} = (i+c-1)^p/(W+(i+c)^p) \approx 1 - (W + p(i+c)^{p-1})/(W+(i+c)^p)$. For $p > 1$ this is approximately $1 - p/i \approx (1-1/i)^p$ for large i, so the indegree distribution again follows a power-law distribution $a_i \sim i^{-p}$. For $p < 1$, however, the constant W dominates the numerator, and the distribution does not follow a power-law.

Model 3: The weight of a vertex w_u is proportional to the PageRank of a vertex. The PageRank (with parameter q) of a vertex is equivalent to the asymptotic fraction of time a surfer that follows a random link from his current location with probability q and jumps to a random vertex with probability $1-q$ spends at a vertex. It is used by search engines as a measure to rank pages; see [2] for more details.

The PageRank model is motivated by the fact that search engines may be introducing feedback into how the Web develops. Users are more likely to link to pages given by search engines, corresponding to pages with high PageRank. Although PageRank is similar to indegree, we expect differences between this model and Model 1, as here new edges have more than a local effect. An alternative and perhaps better model is to have ends of new edges determined by

*Harvard University. Supported by NSF Grant 9983832. edrinea@deas.harvard.edu

†Harvard University. Supported by the Harvard College Research Fund. menaches@fas.harvard.edu.

‡Harvard University. Supported in part by the Sloan Foundation, NSF Grant 9983832, and equipment from Compaq. michaelm@eecs.harvard.edu. Note: this paper was originally submitted in July, 2000, for the SODA 2001 conference.

choosing k vertices uniformly at random and linking to the m vertices with highest PageRank. This potentially mimics the interaction between a user creating a new page and a search engine.

Experiments: We present the results of initial experiments based on these models. More extensive experiments will be given in the full version. Here, for all experiments, each new node is given outdegree 1. All plots are log-log plots with the frequency (in percent) given as a function of the indegree; hence, a straight line plot implies a power-law distribution.

In Figure 1, we examine Model 1 with various c. We present the results from simulating the graph-building process for one million nodes (beginning with 5,000 nodes in a cycle), numerically simulating the corresponding differential equations, and deriving the asymptotic values. We see that the differential equations accurately predict actual behavior, and large graphs are required to reach the asymptotic expressions. (100 million node graphs are much closer to the asymptotic behavior.)

In Figure 2, we examine Model 2 with values $p = 0.8$ and $p = 1.2$. We again build graphs with one million nodes and 5,000 initial nodes and compare to the corresponding differential equations. The equations match the simulations; moreover, the difference in behavior for $p < 1$ and $p > 1$ is evident. We have observed in other experiments that for $p > 1$ it is possible for a single node to take on high degree early in the process, gaining so much weight that it then becomes the endpoint for almost all future edges. This suggests that Model 2 with $p > 1$ is potentially unstable. We hope to examine this point further in future research.

In Figure 3, we examine Model 3, the PageRank model. Here we begin with 50,000 initial nodes in a cycle. Also, we recompute PageRanks only after a batch of new nodes; a batch at time t has $\sqrt{t}$ nodes. The effect of batching appears minor. For large q, too much weight is focused on the initial nodes, causing a hump in the distribution. This suggests that Model 3 may be potentially unstable. For smaller q and a large initial set, the effect disappears.

We expect to expand on these findings in future work.

References

[1] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, vol. 272, pages 173-189, 1999.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of the 7th World Wide Web Conference*, 1998.

[3] 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 *Proc. of the 9th World Wide Web Conference*, 2000.

[4] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Random graph models for the Web graph. To appear in *FOCS 2000*.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large scale knowledge bases from the Web. In *Proc. of the 25th VLDB Conference*, 1999.

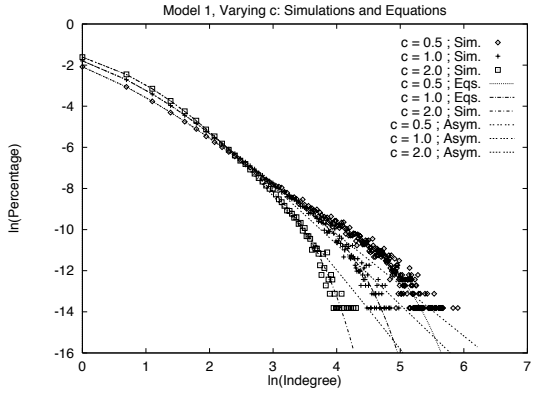

Figure 1: Simulations and results from the differential equations for Model 1.

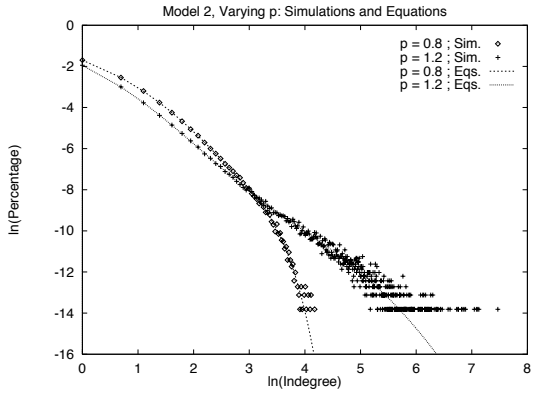

Figure 2: Simulations and results from the differential equations for Model 2, with $c = 1.0$. Different behaviors appear for $p < 1$ and $p > 1$.

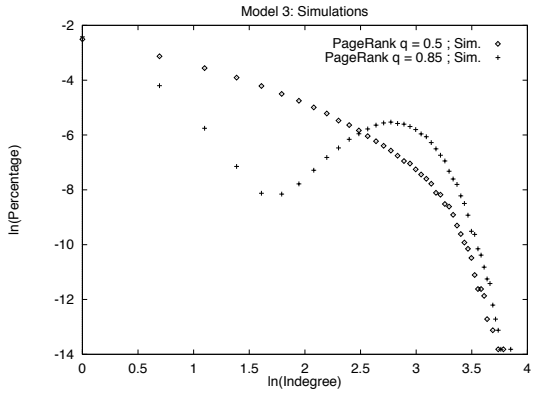

Figure 3: Simulations for Model 3, at $q = 0.5$ and $q = 0.85$.