A Steady State Model for Graph Power Law

David Eppstein* Joseph Wang*

Abstract

Power law distribution seems to be an important characteristic of web graphs. Several existing web graph models [8, 21] generate power law graphs by adding new vertices and non-uniform edge connectivities to existing graphs. Researchers [9, 10, 24] have conjectured that preferential connectivity and incremental growth are both required for the power law distribution. In this paper, we propose a different web graph model with power law distribution that does not require incremental growth. We also provide a comparison of our model with several others in their ability to predict web graph clustering behavior.

1 Introduction

The growth of the World Wide Web (WWW) has been explosive and phenomenal. Google [1] has more than 2 billion pages searched as of February 2002. The Internet Archive [2] has 10 billion pages archived as of March 2001. The existing growth-based models [6, 8, 21] are adequate to explain the web's current graph structure. It would be interesting to know if a different model is needed as the growth rate slows down [3] while its link structure continues to evolve.

1.1 Why Power Laws?

Barabási et al. [9, 10] and Medina et al. [24] stated that preferential connectivity and incremental growth are both required for the power law distribution observed in the web. The importance of the preferential connectivity has been shown by several researchers [8, 16].

Faloutsos et al. [15] observed that the internet topology exhibits power law distribution in the form of $y = x^{\alpha}$. When studying web characteristics, the documents can be viewed as vertices in a graph and the hyper-links as edges between them. Various researchers [7, 8, 19, 22] have independently showed the power law distribution in the degree sequence

*Dept. Inf. & Comp. Sci., UC Irvine, CA 92697-3425, USA, {eppstein,josephw}@ics.uci.edu.

of the web graphs. Huberman and Adamic [5, 16] showed a power law distribution in the web site sizes. [See [20] for a summary of works on the web structure.]

Medina et al. [24] showed that topologies generated by two widely used generators the Waxman model [32], and the GT-ITM tool [13] do not have power law distribution in their degree sequences. Palmer and Steffan [27] proposed a power law degree generator that recursively partitions the adjacency matrix into $80-20$ distribution. However, it is unclear if their generator actually emulates other web properties.

The power law distribution seems to be an ubiquitous property. The power law distribution occurs in epidemics study [30], population study [28], genomes distribution [17, 29], various social phenomena [11, 26], and massive graphs [4, 6]. For the power law graphs in biological systems, the connectivity changes appear to be much more important than growth in size.

1.2 Properties for Graph Model Comparison

Another important property that has been looked at is the diameters of web graphs. However, there are conflicting results in the published papers. Albert et al. [7] stated that the web graphs have the *small world phenomenon* [25, 31], in which the diameter Δ is roughly $0.35 + 2.06 \lg n$, where n is the size of the web graph. For $n = 8 \times 10^8$, $\Delta \approx 19$. Lu [23] proved the diameters of random power law graphs are logarithmic function of n under the model proposed by Aiello et al. [6]. However, Broder et al. [12] showed over 75% of time, there is no directed path between two random vertices. If there is a path, the average distance is roughly 16 when viewing web graph as directed graph or 6.83 in the undirected case.

Currently, there are few theoretical graph models [6, 8, 21, 27] for generating power law graphs. There are very few comparative studies that would allow us to determine which of these theoretical models are more accurate models of the web. We only know that the model proposed by Kumar et al. [21] generates more bipartite cliques than other models. They believe clustering to be an important part of web graph structures that was insufficiently represented in previous models [6, 8].

1.3 New Contributions

In this paper, we show power law graphs do not require incremental growth, by developing a graph model which (empirically) results in power laws by evolving a graph according to a Markov process while maintaining constant size and density. We also provide an easily computable graph property that can be used to capture cluster information in a graph without enumerating all possible subgraphs.

2 Steady State Model

Our *SteadyState* (SS) model is very simple in comparison with other web graph models [6, 8, 21, 27]. It consists of repetitively removing and adding edges on a sparse random graph G.

Let m be $\Theta(n)$. To generate the initial sparse random graph G, we randomly add an edge between vertices with probability $\frac{2m}{n(n-1)}$. If the number of edges in G is still less than m, then we start adding edges between vertices with probability of 0.5 until we have m edges.

We reiterate the following steps r times on G, where r is a parameter to our model.

1. Pick a vertex v at random. If there is no edge incidents upon v, we pick another one.
2. Pick an edge $(u, v) \in G$ at random.
3. Pick a vertex x at random.
4. Pick a vertex y with probability proportional to degree.
5. If (x, y) is not an edge in G and x is not equal to y, then remove edge (u, v) and add edge (x, y).

One can view out model as an aperiodic Markov chain with some limiting distribution. If we repeat the above steps long enough, we will get a random graph drawn from the distribution no matter what the initial random sparse graph is. Note that unlike other models [6, 21], the graphs generated by our model do not contain self-loops nor multiple edges between two vertices.

Barabási et al. [9] also proposed a non-growth model, which failed to produce a power law distribution. Both models have preferential connectivity features. However, there are several differences between our model and theirs. First, our edge set is fixed and the initial graph is generated via classical random graph model [14, 18]. Second, our model has "rewiring" feature similar to one in the small world model [9, 25, 31].

2.1 Simulation Results

We simulated our model on graphs of different sizes, $(500 \leq n \leq 5000)$, and densities $\frac{m}{n}$, $(1 \leq \frac{m}{n} \leq 3)$ for 5 times. We performed $r = 10000000$ edge deletion/insertion operations on each graph. The vertices' degree distributions appear to converge to power law distributions as the number of edge deletion/insertion operations increases. Some of our simulation results are shown in Figures 1 - 4. Figures 1 and 3 show degree distributions at various stage of simulations. Figures 2 and 4 show degree distributions for graphs with different densities

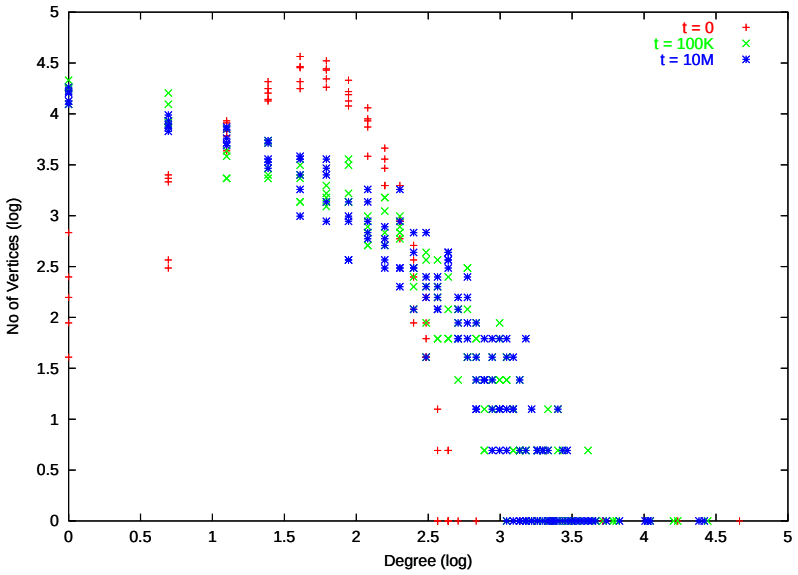

Figure 1: Initial $G(500, 1500)$, & G After $100K$ and $10M$ Steps

$\frac{m}{n}$. (For $G(500, 1500)$, the best lines that fit our log-log plots have slopes between -1.34 and -1.37 and correlation coefficients between 0.808 and 0.877. For $G(3000, 9000)$, the slopes are between -1.51 and -1.62 and correlation coefficients are between 0.76 and 0.81.)

3 Cluster Information

Given a subgraph S of G, $d_S(v)$ is the degree for vertex v in S. Here we examine the maximum degree $d_{\max}$ in all subgraphs, which is defined as

$$d_{\max} = \max_S \ \min_{v \in S} \ d_S(v).$$

We use $d^M_{\max}$ to denote the value obtained under graph model M.

To compute $d_{\max}$ for a graph G, we perform the following steps until G becomes empty:

1. Select a minimum degree vertex v from G.
2. Set $d_{\max}$ to $d(v)$ if $d(v) > d_{\max}$.
3. Remove vertex v and its edges from G.

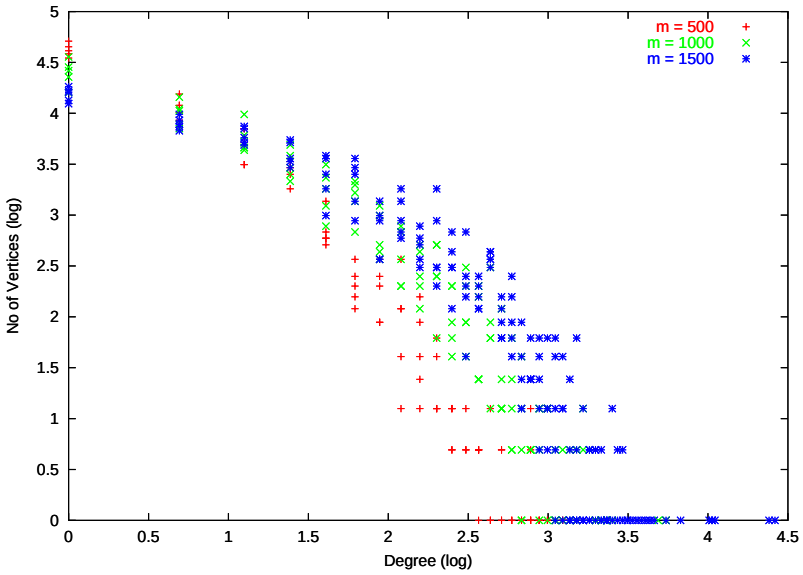

Figure 2: $G(500, 500)$, $G(500, 1000)$, and $G(500, 1500)$ After $10M$ Steps

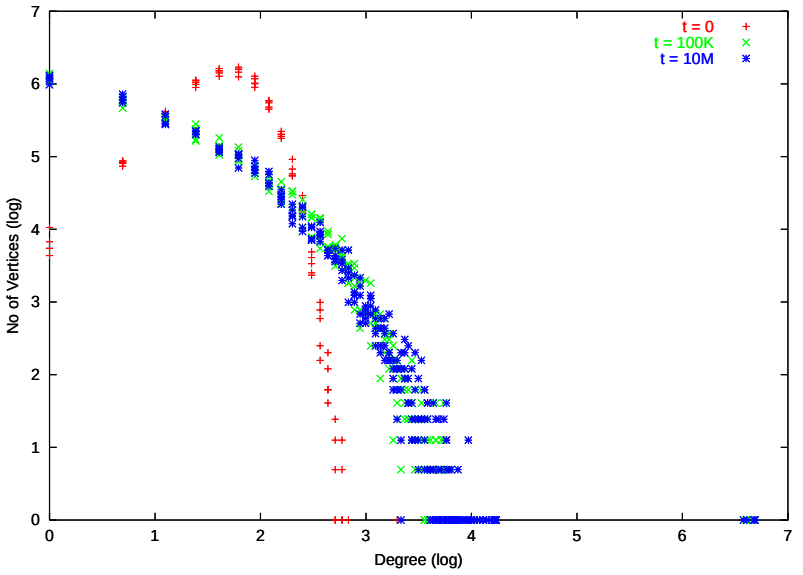

Figure 3: Initial $G(3000, 9000)$, & G After $100K$ and $10M$ Steps

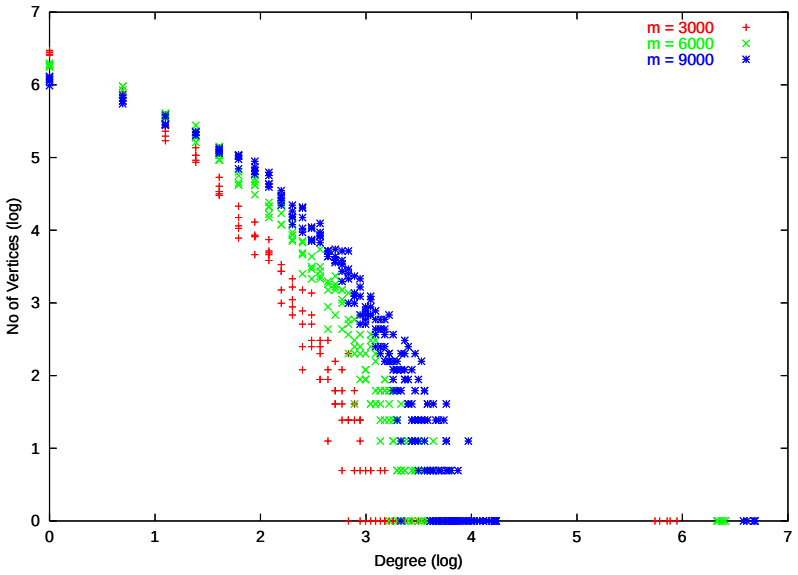

Figure 4: $G(3000, 3000)$, $G(3000, 6000)$, and $G(3000, 9000)$ After $10M$ Steps

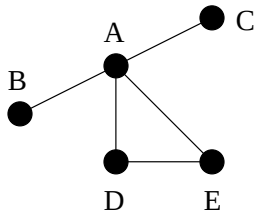

Figure 5: Minimal Degree Vertex Elimination

The above steps correctly compute $d_{\max}$ because we cannot remove any vertices of S until the degree of the current subgraph reaches $d_{\max}$. The minimal degree elimination sequence for graph in Figure 5 will be B, C, A, D, and E. The degrees when those vertices got eliminated are $1, 1, 2, 1$, and 1. $d_{\max}$ is 2 since $\max\{1, 1, 2, 1, 1\} = 2$.

Observation 1 *For any model M that constructs a graph by adding a vertex at a time, and for which each newly added vertex has the same degree $d = \frac{m}{n}$, $d^M_{\max} = d$.*

Thus the Barabási and Albert's model (BA) [8] or the linear growth copying model in [21] has the same value for $d_{\max}$ for graphs of all sizes once $d = \frac{m}{n}$ is fixed.

Observation 2 *The web graph generated by the linear model has minimum vertex degree*

of $d = \frac{m}{n}$.

Hence, the linear model may not encapsulate all the crucial properties in a web graph if there are significant number of vertices with degree less than $\frac{m}{n}$.

3.1 Web Crawl and Simulation Data

We performed web crawl on various Computer Science sites. We then used the ACL model [6] to generate new graphs from degree sequences in the actual web graphs. We also run the SS model using n and m values from the actual web graphs with 10000000 edge insertion/deletion steps. For each graph, we run both models 5 times. The following table shows the means μ and the standard deviations σ for $d_{\max}$ values using the ACL model and the SS model.

Site	n	m	$d_{\max}$	μ_{ACL}	σ_{ACL}	μ_{SS}	σ_{SS}
arizona	5315	16892	15	10	0	8	0
berkeley	2826	22957	45	21.6	0.547	16	0
caltech	622	4830	7	5.8	0.447	12.8	0.447
cmu	2052	23821	57	37.2	0.447	20	0.707
cornell	7145	14919	17	19.4	0.547	6	0
harvard	915	9327	21	12.6	0.894	16.4	0.547
mit	4861	15360	31	24.4	0.547	7	0
nd	1913	16328	33	29.2	0.447	15.4	0.547
stanford	2553	25693	27	14.6	0.547	18.4	0.547
ucla	2718	19755	22	16.6	0.547	14.2	0.447
ucsb	5236	10338	22	13.8	0.447	5	0
ucsd	553	3885	15	7.2	0.447	11.8	0.447
uiowa	1410	12258	8	8.8	0.447	15.2	0.447
uiuc	5623	28872	29	21	0	11.8	0.836
unc	1465	5446	17	9.8	0.447	8	0
washington	7001	24901	17	12	0	9	0

Table 1: $d_{\max}$ from Actual Web Crawl and Models Simulation

In general, the ACL model and the SS model are generating less clustered graphs than what we see on actual web graphs. This implies that we need a more detailed model of web graph clustering behavior.

4 Conclusion and Open Problems

Previously, researchers have conjectured that preferential connectivity and incremental growth are necessary factors in creating power law graphs. In this paper, we provide a model of graph evolution that produces power law without growth. Our *SteadyState* model is very simple in comparison with other graph models [21]. It also does not require prior degree sequences as in the *ACL* model [6].

The difficulty in comparing various models [6, 8, 21] is that each model has different parameters and inputs. Here we provide a simple graph property d_{max} that captures the clustering behavior of graphs without complicated subgraph enumeration algorithm. It can be useful in gauging the accuracy of various models.

From our web crawl data, we know that the linear models such as Barabási's [8] are not the best ones to use when considering $d_{\max}$. Both *ACL* and *SS* models are not generating dense-enough subgraphs when comparing against the actual web graphs. Thus, we need a better web graph model that mimics actual web graph clustering behavior.

Here are some of our open problems:

1. Can one prove theoretically that the SS method actually has a power law distribution?

2. How long does it take for our model to reach a steady state? As time proceeds, the "high" degree vertices will attract more edges whereas all other vertices will have fewer edges connecting to them until we reach a state, after which the degree distribution won't fluctuate much.

3. What are other simple web graph properties that we can use to determine the accuracy of various models?

4. Are there any technique such as graph product that we can use to generate massive web graphs in relative short time?

References

[1] Google. www.google.com.

[2] The Internet Archive. www.archive.org.

[3] Online Computer Library Center. wcp.oclc.org.

[4] Abello, J., Buchsbaum, A., and Westbrook, J. A functional approach to external graph algorithms. In *Proceedings of 6th Eurpoean Syposium on Algorithms* (1998), pp. 332–343.

[5] Adamic, L., and Huberman, B. Power-law distribution of the world wide web. *Science 287* (2000), 2115.

[6] Aiello, W., Chung, F., and Lu, L. A random graph model for massive graphs. In *Proceedings on Theory of Computing* (2000), pp. 171–180.

[7] Albert, R., Jeong, H., and Barabási, A. Diameter of the world-wide web. *Nature 401* (September 1999), 130–131.

[8] Barabási, A., and Albert, R. Emergence of scaling in random networks. *Sicence 286*, 5439 (1999), 509–512.

[9] Barabási, A., Albert, R., and Jeong, H. Mean-field theory for scale-free random networks. *Physica A 272* (1999), 173–187.

[10] Barabási, A., Albert, R., and Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A 281* (2000), 69–77.

[11] Barabási, A., Albert, R., Jeong, H., and Bianconi, G. Pow-law distribution of the world wide web. *Science 287* (2000), 2115.

[12] Broder, A. Z., Kumar, S. R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the web: experiements and models. In *Proceedings of 9th WWW Conference* (2000), pp. 309–320.

[13] Calvert, K., Doar, M., and Zegura, E. Modeling internet topology. *IEEE Communications Magazine* (June 1997), 160–163.

[14] Erdős, P., and Rényi, A. On random graphs i. *Publ. Math. Dececen 6* (1959), 290–297.

[15] Faloutsos, M., Faloutsos, P., and Faloutsos, C. On power-law relationship of the internet topology. In *Proceedings of the ACM SIGCOM Conference* (1999), pp. 251–260.

[16] Huberman, B., and Adamic, L. Growth dynamics of the world-Wide web. *Science 401* (September 1999), 131–131.

[17] Huynen, M. A., and van Nimwegen, E. Power laws in the size distribution of gene families in complete genomes: biological interpretations. Tech. Rep. 97-03-025, Santa Fe Institue, 1997.

[18] Janson, S., Luczak, T., and Rucinski, A. *Random Graphs.* John Wiley & Sons, 2000.

[19] Kleinberg, J., Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A. The web as a graph: Measurements, models and methods. In *Proceedings on Combinatorics and Computing* (1999), pp. 1–18.

[20] Kleinberg, J., and Lawrence, S. The structure of the web. *Science 294* (2001), 1849–1850.

[21] Kumar, S. R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. Stochastic models for the web graph. In *Proceedings on Foundations of Computer Science* (2000), pp. 57–65.

[22] Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Trawling the weeb for emerging cyber-communities. In *Proceedings of 8th WWW Conference* (1999), pp. 403–416.

[23] Lu, L. The diameter of random massive graphs. In *Proceedings on Discrete Algorithms* (2001), pp. 912–921.

[24] Medina, A., Matta, I., and Byers, J. On the origin of power laws in internet topologies. *ACM Computer Communication Review 30*, 2 (2000), 18–28.

[25] Milgram, S. The small world problem. *Psychol. Today 2* (1967), 60–67.

[26] Ormerod, P., and Smith, L. Power law distribution of lifespans of large firms: breakdown of scaling. Tech. rep., Volterra Consulting Ltd., 2001.

[27] Palmer, C., and Steffan, J. Generating network topologies that obey power laws. In *Proceedings of IEEE Globecom* (2000).

[28] Palmer, M. W., and Whitge, P. S. Scale dependence and the species-area relationship. *American Naturalist 144* (1994), 717–740.

[29] Qian, J., Luscombe, N. M., and Gerstein, M. Proten family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Mol. Biology 313* (2001), 673–681.

[30] Rhodes, C. J., and Anderson, R. M. Power laws governing epidemics in isolated popluations. *Nature 381* (1996), 600–602.

[31] Watts, D. J. *Small worlds: the dynamics of networks between order and randomness.* Princeton University Press, Princeton, N.J., 1999.

[32] Waxman, B. M. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communication 6*, 9 (December 1988), 1617–1622.