Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations

Jure Leskovec
Carnegie Mellon University
jure@cs.cmu.edu

Jon Kleinberg*
Cornell University
kleinber@cs.cornell.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

ABSTRACT

How do real graphs evolve over time? What are "normal" growth patterns in social, technological, and information networks? Many studies have discovered patterns in *static graphs*, identifying properties in a single snapshot of a large network, or in a very small number of snapshots; these include heavy tails for in- and out-degree distributions, communities, small-world phenomena, and others. However, given the lack of information about network evolution over long periods, it has been hard to convert these findings into statements about trends over time.

Here we study a wide range of real graphs, and we observe some surprising phenomena. First, most of these graphs densify over time, with the number of edges growing superlinearly in the number of nodes. Second, the average distance between nodes often *shrinks* over time, in contrast to the conventional wisdom that such distance parameters should increase slowly as a function of the number of nodes (like $O(\log n)$ or $O(\log(\log n))$.

Existing graph generation models do not exhibit these types of behavior, even at a qualitative level. We provide a new graph generator, based on a "forest fire" spreading process, that has a simple, intuitive justification, requires very few parameters (like the "flammability" of nodes), and produces graphs exhibiting the full range of properties observed both in prior work and in the present study.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – Data Mining

General Terms

Measurement, Theory

Keywords

densification power laws, graph generators, graph mining, heavy-tailed distributions, small-world phenomena

1. INTRODUCTION

In recent years, there has been considerable interest in graph structures arising in technological, sociological, and scientific settings: computer networks (routers or autonomous systems connected together); networks of users exchanging e-mail or instant messages; citation networks and hyperlink networks; social networks (who-trusts-whom, who-talks-to-whom, and so forth); and countless more [24]. The study of such networks has proceeded along two related tracks: the measurement of large network datasets, and the development of random graph models that approximate the observed properties.

Many of the properties of interest in these studies are based on two fundamental parameters: the nodes' *degrees* (i.e., the number of edges incident to each node), and the *distances* between pairs of nodes (as measured by shortest-path length). The node-to-node distances are often studied in terms of the *diameter* — the maximum distance — and a set of closely related but more robust quantities including the average distance among pairs and the *effective diameter* (the 90th percentile distance, a smoothed form of which we use for our studies).

Almost all large real-world networks evolve over time by the addition and deletion of nodes and edges. Most of the recent models of network evolution capture the growth process in a way that incorporates two pieces of "conventional wisdom:"

(A) *Constant average degree assumption*: The average node degree in the network remains constant over time. (Or equivalently, the number of edges grows linearly in the number of nodes.)

(B) *Slowly growing diameter assumption*: The diameter is a slowly growing function of the network size, as in "small world" graphs [4, 7, 22, 30].

*This research was done while on sabbatical leave at CMU.

Work partially supported by the National Science Foundation under Grants No. IIS-0209107, SENSOR-0329549, IIS-0326322, CNS-0433540, CCF-0325453, IIS-0329064, CNS-0403340, CCR-0122581, a David and Lucile Packard Foundation Fellowship, and also by the Pennsylvania Infrastructure Technology Alliance (PITA), a partnership of Carnegie Mellon, Lehigh University and the Commonwealth of Pennsylvania's Department of Community and Economic Development (DCED). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

2. RELATED WORK

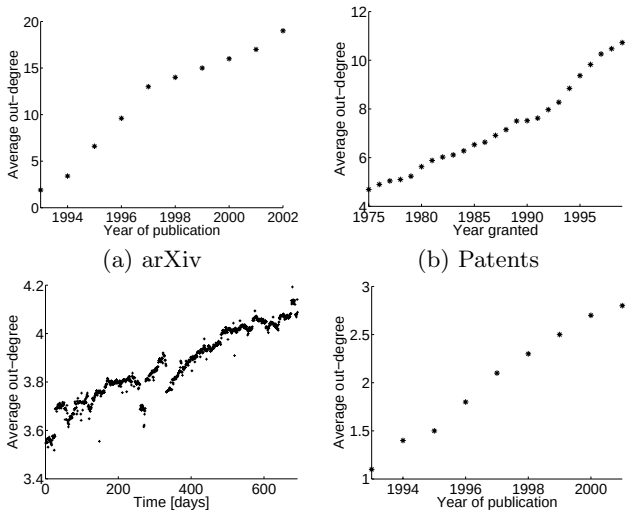
Research over the past few years has identified classes of properties that many real-world networks obey. One of the main areas of focus has been on *degree power laws*, showing that the set of node degrees has a heavy-tailed distribution. Such degree distributions have been identified in phone call graphs [1], the Internet [11], the Web [3, 14, 20], click-stream data [5] and for a who-trusts-whom social network [8]. Other properties include the "small-world phenomenon," popularly known as "six degrees of separation", which states that real graphs have surprisingly small (average or effective) diameter (see [4, 6, 7, 9, 17, 22, 30, 31]).

In parallel with empirical studies of large networks, there has been considerable work on probabilistic models for graph generation. The discovery of degree power laws led to the development of random graph models that exhibited such degree distributions, including the family of models based on *preferential attachment* [2, 3, 10] and the related *copying model* [18, 19]. See [23, 24] for surveys of this area.

It is important to note the fundamental contrast between one of our main findings here — that the average number of out-links per node is growing polynomially in the network size — and body of work on degree power laws. This earlier work developed models that almost exclusively used the assumption of node degrees that were bounded by constants (or at most logarithmic functions) as the network grew; our findings and associated model challenge this assumption, by showing that networks across a number of domains are becoming *denser*.

The bulk of prior work on the study of network datasets has focused on *static* graphs, identifying patterns in a single snapshot, or a small number of network snapshots (see also the discussion of this point by Ntoulas et al. [25]). Two exceptions are the very recent work of Katz [16], who independently discovered densification power laws for citation networks, and the work of Redner [28], who studied the evolution of the citation graph of *Physical Review* over the past century. Katz's work builds on his earlier research on power-law relationships between the size and recognition of professional communities [15]; his work on densification is focused specifically on citations, and he does not propose a generative network model to account for the densification phenomenon, as we do here. Redner's work focuses on a range of citation patterns over time that are different from the network properties we study here.

Our Community Guided Attachment (CGA) model, which produces densifying graphs, is an example of a hierarchical graph generation model, in which the linkage probability between nodes decreases as a function of their relative distance in the hierarchy [8, 17, 31]. Again, there is a distinction between the aims of this past work and our model here; where these earlier network models were seeking to capture properties of individual snapshots of a graph, we seek to explain a time evolution process in which one of the fundamental parameters, the average node degree, is varying as the process unfolds. Our Forest Fire Model follows the overall framework of earlier graph models in which nodes arrive one at a time and link into the existing structure; like the copying model discussed above, for example, a new node creates links by consulting the links of existing nodes. However, the recursive process by which nodes in the Forest Fire Model creates these links is quite different, leading to the new properties discussed in the previous section.



(a) arXiv (b) Patents

(c) Autonomous Systems (d) Affiliation network

Figure 1: The average node out-degree over time. Notice that it increases, in all 4 datasets. That is, all graphs are *densifying*.

3. OBSERVATIONS

We study the temporal evolution of several networks, by observing snapshots of these networks taken at regularly spaced points in time. We use datasets from four different sources; for each, we have information about the time when each node was added to the network over a period of several years — this enables the construction of a snapshot at any desired point in time. For each of datasets, we find a version of the densification power law from Equation (1), $e(t) \propto n(t)^a$; the exponent a differs across datasets, but remains remarkably stable over time. We also find that the effective diameter decreases in all the datasets considered.

The datasets consist of two citation graphs for different areas in the physics literature, a citation graph for U.S. patents, a graph of the Internet, and five bipartite affiliation graphs of authors with papers they authored. Overall, then, we consider 9 different datasets from 4 different sources.

3.1 Densification Laws

Here we describe the datasets we used, and our findings related to densification. For each graph dataset, we have, or can generate, several time snapshots, for which we study the number of nodes $n(t)$ and the number of edges $e(t)$ at each timestamp t. We denote by n and e the final number of nodes and edges. We use the term *Densification Power Law plot* (or just DPL plot) to refer to the log-log plot of number of edges $e(t)$ versus number of nodes $n(t)$.

3.1.1 ArXiv citation graph

We first investigate a citation graph provided as part of the 2003 KDD Cup [12]. The HEP–TH (high energy physics theory) citation graph from the e-print arXiv covers all the citations within a dataset of n=29,555 papers with e= 352,807 edges. If a paper i cites paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this. We refer to this dataset as *arXiv*.

son, the date of their first submission to the arXiv. The data for affiliation graphs covers the period from April 1992 to March 2002. The smallest of the graphs (category GR–QC) had 19,309 nodes (5,855 authors, 13,454 papers) and 26,169 edges. ASTRO–PH is the largest graph, with 57,381 nodes (19,393 authors, 37,988 papers) and 133,170 edges. It has 6.87 authors per paper; most of the other categories also have similarly high numbers of authors per paper.

For all these affiliation graphs we observe similar phenomena, and in particular we have densification exponents between 1.08 and 1.15. Due to lack of space we present the complete set of measurements only for ASTRO–PH, the largest affiliation graph. Figures 1(d) and 2(d) show the increasing average degree over time, and a densification exponent of $a = 1.15$.

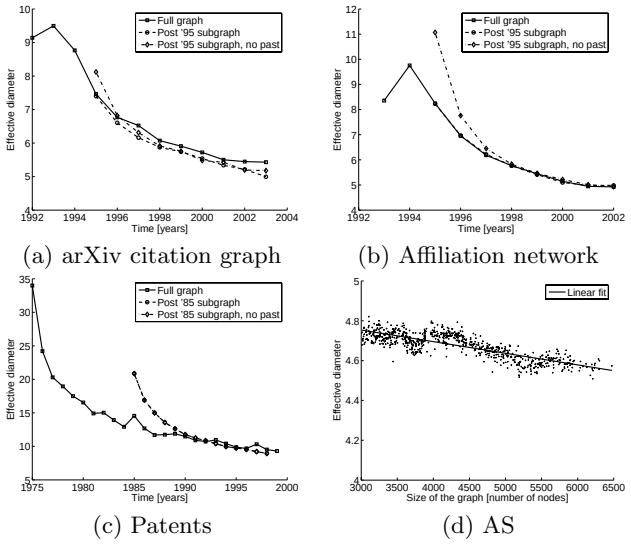
3.2 Shrinking Diameters

We now discuss the behavior of the effective diameter over time, for this collection of network datasets. Following the conventional wisdom on this topic, we expected the underlying question to be whether we could detect the differences among competing hypotheses concerning the growth rates of the diameter — for example, the difference between logarithmic and sub-logarithmic growth. Thus, it was with some surprise that we found the effective diameters to be actually *decreasing* over time (Figure 3).

Let us make the definitions underlying the observations concrete. We say that two nodes in an undirected network are *connected* if there is an path between them; for each natural number d, let $g(d)$ denote the fraction of connected node pairs whose shortest connecting path has length at most d. The *hop-plot* for the network is the set of pairs $(d, g(d))$; it thus gives the cumulative distribution of distances between connected node pairs. We extend the hop-plot to a function defined over all positive real numbers by linearly interpolating between the points $(d, g(d))$ and $(d+1, g(d+1))$ for each d, and we define the *effective diameter* of the network to be the value of d at which this function achieves the value 0.9. (Note that this varies slightly from an alternate definition of the effective diameter used in earlier work: the minimum value d such that at least 90% of the connected node pairs are at distance at most d. Our variation smooths this definition by allowing it to take non-integer values.) The effective diameter is a more robust quantity than the diameter (defined as the maximum distance over all connected node pairs), since the diameter is prone to the effects of degenerate structures in the graph (e.g. very long chains). However, the effective diameter and diameter tend to exhibit qualitatively similar behavior.

For each time t (as in the previous subsection), we create a graph consisting of nodes up to that time, and compute the effective diameter of the undirected version of the graph.

Figure 3 shows the effective diameter over time; one observes a decreasing trend for all the graphs. We performed a comparable analysis to what we describe here for all 9 graph datasets in our study, with very similar results. For the citation networks in our study, the decreasing effective diameter has the following interpretation: Since all the links out of a node are "frozen" at the moment it joins the graph, the decreasing distance between pairs of nodes appears to be the result of subsequent papers acting as "bridges" by citing earlier papers from disparate areas. Note that for other graphs in our study, such as the AS dataset, it is possible for



(a) arXiv citation graph (b) Affiliation network

(c) Patents (d) AS

Figure 3: The effective diameter over time.

an edge between two nodes to appear at an arbitrary time after these two nodes join the graph.

We note that the effective diameter of a graph over time is necessarily bounded from below, and the decreasing patterns of the effective diameter in the plots of Figure 3 are consistent with convergence to some asymptotic value. However, understanding the full "limiting behavior" of the effective diameter over time, to the extent that this is even a well-defined notion, remains an open question.

3.2.1 Validating the shrinking diameter conclusion

Given the unexpected nature of this result, we wanted to verify that the shrinking diameters were not attributable to artifacts of our datasets or analyses. We explored this issue in a number of ways, which we now summarize; the conclusion is that the shrinking diameter appears to be a robust, and intrinsic, phenomenon. Specifically, we performed experiments to account for (a) possible sampling problems, (b) the effect of disconnected components, (c) the effect of the "missing past" (as in the previous subsection), and (d) the dynamics of the emergence of the giant component.

Possible sampling problems: Computing shortest paths among all node pairs is computationally prohibitive for graphs of our scale. We used several different approximate methods, obtaining almost identical results from all of them. In particular, we applied the Approximate Neighborhood Function (ANF) approach [27] (in two different implementations), which can estimate effective diameters for very large graphs, as well as a basic sampling approach in which we ran exhaustive breadth-first search from a subset of the nodes chosen uniformly at random. The results using all these methods were essentially identical.

Disconnected components: One can also ask about the effect of small disconnected components. All of our graphs have a single *giant component* – a connected component (or weakly connected component in the case of directed graphs, ignoring the direction of the edges) that accounts for a significant fraction of all nodes. For each graph, we computed effective diameters for both the entire graph and for just the

Table 1: Table of symbols

Symbol	Description
a	*Densification Exponent*
c	*Difficulty Constant*
$f(h)$	*Difficulty Function*
$n(t)$	number of nodes at time t
$e(t)$	number of edges at time t
b	community branching factor
$\bar{d}$	expected average node out-degree
H	height of the tree
$h(v,w)$	least common ancestor height of v, w
p	forest fire forward burning probability
p_b	forest fire backward burning probability
r	ratio of backward and forward probability

We take the following approach. Power laws often appear in combination with *self-similar* datasets [29]. Our approach involves two steps, both of which are based on self-similarity. Thus, we begin by searching for self-similar, recursive structures. In fact, we can easily find several such recursive sets: For example, computer networks form tight groups (e.g., based on geography), which consist of smaller groups, and so on, recursively. Similarly for patents: they also form conceptual groups ("chemistry", "communications", etc.), which consist of sub-groups, and so on recursively. Several other graphs feature such "communities within communities" patterns. For example, it has been argued (see e.g. [31] and the references therein) that social structures exhibit self-similarity, with individuals organizing their social contacts hierarchically. Moreover, pairs of individuals belonging to the same small community form social ties more easily than pairs of individuals who are only related by membership in a larger community. In a different domain, Menczer studied the frequency of links among Web pages that are organized into a topic hierarchy such as the Open Directory [21]. He showed that link density among pages decreases with the height of their least common ancestor in the hierarchy. That is, two pages on closely related topics are more likely to be hyperlinked than are two pages on more distantly related topics.

This is the first, qualitative step in our explanation for the Densification Power Law. The second step is quantitative. We will need a numerical measure of the difficulty in crossing communities; we call this the *Difficulty Constant*, and we define it more precisely below.

4.1.1 The Basic Version of the Model

We represent the recursive structure of communities-within-communities as a tree Γ, of height H. We shall show that even a simple, perfectly balanced tree of constant fanout b is enough to lead to a densification power law, and so we will focus the analysis on this basic model.

The nodes V in the graph we construct will be the leaves of the tree; that is, $n = |V|$. (Note that $n = b^H$.) Let $h(v,w)$ define the standard tree distance of two leaf nodes v and w: that is, $h(v,w)$ is the height of their least common ancestor (the height of the smallest sub-tree containing both v and w).

We will construct a random graph on a set of nodes V by specifying the probability that v and w form a link as a function f of $h(v,w)$. We refer to this function f as the *Difficulty Function*. What should be the form of f? Clearly, it should decrease with h; but there are many forms such a decrease could take.

The form of f that works best for our purposes comes from the self-similarity arguments we made earlier: We would like f to be scale-free; that is, $f(h)/f(h-1)$ should be level-independent and thus constant. The only way to achieve level-independence is to define $f(h) = f(0) * c^{-h}$. Setting $f(0)$ to 1 for simplicity, we have:

$$f(h) = c^{-h} \tag{2}$$

where $c \geq 1$. We refer to the constant c as the *Difficulty Constant*. Intuitively, cross-communities links become harder to form as c increases.

This completes our development of the model, which we refer to as *Community Guided Attachment*: If the nodes of a graph belong to communities-within-communities, and if the cost for cross-community edges is scale-free (Eq. (2)), the Densification Power Law follows naturally. No central control or exogenous regulations are needed to force the resulting graph to obey this property. In short, self-similarity itself leads to the Densification Power Law.

THEOREM 1. *In the Community Guided Attachment random graph model just defined, the expected average out-degree $\bar{d}$ of a node is proportional to:*

$$\begin{aligned} \bar{d} &= n^{1-\log_b(c)} \quad \text{if } 1 \leq c < b & (3)\\ &= \log_b(n) \quad \text{if } c = b & (4)\\ &= \text{constant} \quad \text{if } c > b & (5) \end{aligned}$$

PROOF. For a given node v, the expected out-degree (number of links) $\bar{d}$ of the node is proportional to

$$\bar{d} = \sum_{x \neq v} f(h(x,v)) = \sum_{j=1}^{\log_b(n)} (b-1) b^{j-1} c^{-j} = \frac{b-1}{c} \sum_{j=1}^{\log_b(n)} \left(\frac{b}{c}\right)^{j-1}. \tag{6}$$

There are three different cases: if $1 \leq c < b$ then by summing the geometric series we obtain

$$\begin{aligned} \bar{d} &= \frac{b-1}{c} \cdot \frac{\left(\frac{b}{c}\right)^{\log_b(n)} - 1}{\left(\frac{b}{c}\right) - 1} = \left(\frac{b-1}{b-c}\right)\left(n^{1-\log_b(c)} - 1\right)\\ &= \Theta(n^{1-\log_b(c)}). \end{aligned}$$

In the case when $c = b$ the series sums to

$$\begin{aligned} \bar{d} &= \sum_{x \neq v} f(h(x,v)) = \frac{b-1}{b} \sum_{j=1}^{\log_b(n)} \left(\frac{b}{b}\right)^{j-1} = \frac{b-1}{b} \log_b(n)\\ &= \Theta(\log_b(n)). \end{aligned}$$

The last case is when Difficulty Constant c is greater than branching factor b ($c > b$), then the sum in Eq. (6) converges to a constant even if carried out to infinity, and so we obtain $\bar{d} = \Theta(1)$. □

Note that when $c < b$, we get a densification law with exponent greater than 1: the expected out-degree is $n^{1-\log_b(c)}$, and so the total number of edges grows as n^a where $a = 2 - \log_b(c)$. Moreover, as c varies over the interval $[1, b)$, the exponent a ranges over all values in the interval $(1, 2]$.

4.2.1 The Basic Forest Fire Model

Following this plan, we now define the most basic version of the model. Essentially, nodes arrive one at a time and form out-links to some subset of the earlier nodes; to form out-links, a new node v attaches to a node w in the existing graph, and then begins "burning" links outward from w, linking with a certain probability to any new node it discovers. One can view such a process as intuitively corresponding to a model by which an author of a paper identifies references to include in the bibliography. He or she finds a first paper to cite, chases a subset of the references in this paper (modeled here as random), and continues recursively with the papers discovered in this way. Depending on the bibliographic aids being used in this process, it may also be possible to chase back-links to papers that cite the paper under consideration. Similar scenarios can be considered for social networks: a new computer science graduate student arrives at a university, meets some older CS students, who introduce him/her to their friends (CS or non-CS), and the introductions may continue recursively.

We formalize this process as follows, obtaining the Forest Fire Model. To begin with, we will need two parameters, a *forward burning probability* p, and a *backward burning ratio* r, whose roles will be described below. Consider a node v joining the network at time $t > 1$, and let G_t be the graph constructed thus far. (G_1 will consist of just a single node.) Node v forms out-links to nodes in G_t according to the following process.

(i) v first chooses an *ambassador node* w uniformly at random, and forms a link to w.

(ii) We generate a random number x that is binomially distributed with mean $(1-p)^{-1}$. Node v selects x links incident to w, choosing from among both out-links and in-links, but selecting in-links with probability r times less than out-links. Let $w_1, w_2, \ldots, w_x$ denote the other ends of these selected links.

(iii) v forms out-links to $w_1, w_2, \ldots, w_x$, and then applies step (ii) recursively to each of $w_1, w_2, \ldots, w_x$. As the process continues, nodes cannot be visited a second time, preventing the construction from cycling.

Thus, the "burning" of links in Forest Fire model begins at w, spreads to $w_1, \ldots, w_x$, and proceeds recursively until it dies out. In terms of the intuition from citations in papers, the author of a new paper v initially consults w, follows a subset of its references (potentially both forward and backward) to the papers $w_1, \ldots, w_x$, and then continues accumulating references recursively by consulting these papers. The key property of this model is that certain nodes produce large "conflagrations," burning many edges and hence forming many out-links before the process ends.

Despite the fact that there is no explicit hierarchy in the Forest Fire Model, as there was in Community Guided Attachment, there are some subtle similarities between the models. Where a node in Community Guided Attachment was the child of a parent in the hierarchy, a node v in the Forest Fire Model also has an "entry point" via its chosen ambassador node w. Moreover, just as the probability of linking to a node in Community Guided Attachment decreased exponentially in the tree distance, the probability that a new node v burns k successive links so as to reach a node u lying k steps away is exponentially small in k. (Of course, in the Forest Fire Model, there may be many paths that could be burned from v to u, adding some complexity to this analogy.)

In fact, our Forest Fire Model combines the flavors of several older models, and produces graphs qualitatively matching their properties. We establish this by simulation, as we describe below, but it is also useful to provide some intuition for why these properties arise.

• *Heavy-tailed in-degrees.* Our model has a "rich get richer" flavor: highly linked nodes can easily be reached by a newcomer, no matter which ambassador it starts from.

• *Communities.* The model also has a "copying" flavor: a newcomer copies several of the neighbors of his/her ambassador (and then continues this recursively).

• *Heavy-tailed out-degrees.* The recursive nature of link formation provides a reasonable chance for a new node to burn many edges, and thus produce a large out-degree.

• *Densification Power Law.* A newcomer will have a lot of links near the community of his/her ambassador; a few links beyond this, and significantly fewer farther away. Intuitively, this is analogous to the Community Guided Attachment, although without an explicit set of communities.

• *Shrinking diameter.* It is not a priori clear why the Forest Fire Model should exhibit a shrinking diameter as it grows. Graph densification is helpful in reducing the diameter, but it is important to note that densification is certainly not enough on its own to imply shrinking diameter. For example, the Community Guided Attachment model obeys the Densification Power Law, but it can be shown to have a diameter that slowly increases.

Rigorous analysis of the Forest Fire Model appears to be quite difficult. However, in simulations, we find that by varying just the two parameters p and r, we can produce graphs that densify ($a > 1$), exhibit heavy-tailed distributions for both in- and out-degrees (Fig. 6), and have diameters that decrease. This is illustrated in Figure 5, which shows plots for the effective diameter and the Densification Power Law exponent as a function of time for some selections of p and r. We see from these plots that, depending on the forward and backward burning parameters, the Forest Fire Model is capable of generating sparse or dense graphs, with effective diameters that either increase or decrease.

4.2.2 Extensions to the Forest Fire Model

Our basic version of the Forest Fire Model exhibits rich structure with just two parameters. By extending the model in natural ways, we can fit observed network data even more closely. We propose two natural extensions: "*orphans*" and multiple ambassadors.

"*Orphans*": In both the patent and arXiv citation graphs, there are many isolated nodes, that is, documents with no citations into the corpus. For example, many papers in the arXiv only cite non-arXiv papers. We refer to them as *orphans*. Our basic model does not produce orphans, since each node always links at least to its chosen ambassador. However, it is easy to incorporate orphans into the model in two different ways. We can start our graphs with $n_0 > 1$ nodes at time $t = 1$; or we can have some probability $q > 0$ that a newcomer will form no links (not even to its ambassador).

We find that such variants of the model have a more pronounced decrease in the effective diameter over time, with

• The Densification Power Law: In contrast to the standard modeling assumption that the average out-degree remains constant over time, we discover that real graphs have out-degrees that grow over time, following a natural pattern (Eq. (1)).

• Shrinking diameters: Our experiments also show that the standard assumption of slowly growing diameters does not hold in a range of real networks; rather, the diameter may actually exhibit a gradual decrease as the network grows.

• We show that our Community Guided Attachment-model can lead to the Densification Power Law, and that it needs only one parameter to achieve it.

• Finally, we give the Forest Fire Model, based on only two parameters, which is able to capture patterns observed both in previous work and in the current study: heavy-tailed in- and out-degrees, the Densification Power Law, and a shrinking diameter.

Our results have potential relevance in multiple settings, including 'what if' scenarios; in forecasting of future parameters of computer and social networks; in anomaly detection on monitored graphs; in designing graph sampling algorithms; and in realistic graph generators.

Acknowledgements: We would like to thank Michalis Faloutsos and George Siganos of UCR, for help with the data and for early discussions on the Autonomous System dataset.

6. REFERENCES

[1] J. Abello, A. L. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. In *Proceedings of the 6th Annual European Symposium on Algorithms*, pages 332–343. Springer-Verlag, 1998.

[2] J. Abello, P. M. Pardalos, and M. G. C. Resende. *Handbook of massive data sets*. Kluwer, 2002.

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

[4] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world-wide web. *Nature*, 401:130–131, September 1999.

[5] Z. Bi, C. Faloutsos, and F. Korn. The dgx distribution for mining massive, skewed data. In *KDD*, pages 17–26, 2001.

[6] B. Bollobas and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1), 2004.

[7] 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 World Wide Web Conference*, 2000.

[8] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, 2004.

[9] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[10] C. Cooper and A. Frieze. A general model of web graphs. *Random Struct. Algorithms*, 22(3):311–335, 2003.

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

[12] J. Gehrke, P. Ginsparg, and J. M. Kleinberg. Overview of the 2003 kdd cup. *SIGKDD Explorations*, 5(2):149–151, 2003.

[13] B. H. Hall, A. B. Jaffe, and M. Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. NBER Working Papers 8498, National Bureau of Economic Research, Inc, Oct. 2001.

[14] B. A. Huberman and L. A. Adamic. Growth dynamics of the world-wide web. *Nature*, 399:131, 1999.

[15] J. S. Katz. The self-similar science system. *Research Policy*, 28:501–517, 1999.

[16] J. S. Katz. Scale independent bibliometric indicators. *Measurement: Interdisciplinary Research and Perspectives*, 3:24–28, 2005.

[17] J. M. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems 14*, 2002.

[18] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. In *Proc. International Conference on Combinatorics and Computing*, pages 1–17, 1999.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symp. on Foundations of Computer Science*, 2000.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of 8th International World Wide Web Conference*, 1999.

[21] F. Menczer. Growing and navigating the small world web by local content. *Proceedings of the National Academy of Sciences*, 99(22):14014–14019, 2002.

[22] S. Milgram. The small-world problem. *Psychology Today*, 2:60–67, 1967.

[23] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions, 2004.

[24] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[25] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. In *WWW Conference*, pages 1–12, New York, New York, May 2004.

[26] U. of Oregon Route Views Project. Online data and reports. http://www.routeviews.org.

[27] C. R. Palmer, P. B. Gibbons, and C. Faloutsos. Anf: A fast and scalable tool for data mining in massive graphs. In *SIGKDD*, Edmonton, AB, Canada, 2002.

[28] S. Redner. Citation statistics from more than a century of physical review. Technical Report physics/0407137, arXiv, 2004.

[29] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman and Company, New York, 1991.

[30] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[31] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.