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.

For example, the intensively-studied *preferential attachment model* [3, 24] posits a network in which each new node, when it arrives, attaches to the existing network by a constant number of out-links, according to a "rich-get-richer" rule. Recent work has given tight asymptotic bounds on the diameter of preferential attachment networks [6, 9]; depending on the precise model, these bounds grow logarithmically or even slower than logarithmically in the number of nodes.

How are assumptions (A) and (B) reflected in data on network growth? Empirical studies of large networks to date have mainly focused on *static* graphs, identifying properties of a single snapshot or a very small number of snapshots of a large network. For example, despite the intense interest in the Web's link structure, the recent work of Ntoulas et al. [25] noted the lack of prior empirical research on the evolution of the Web. Thus, while one can assert based on these studies that, qualitatively, real networks have relatively small average node degrees and diameters, it has not been clear how to convert these into statements about trends over time.

The present work: Densification laws and shrinking diameters. Here we study a range of different networks, from several domains, and we focus specifically on the way in which fundamental network properties vary with time. We find, based on the growth patterns of these networks, that principles (A) and (B) need to be reassessed. Specifically, we show the following for a broad range of networks across diverse domains.

(A') *Empirical observation: Densification power laws*: The networks are becoming *denser* over time, with the average degree increasing (and hence with the number of edges growing super-linearly in the number of nodes). Moreover, the densification follows a power-law pattern.

(B') *Empirical observation: Shrinking diameters:* The effective diameter is, in many cases, actually *decreasing* as the network grows.

We view the second of these findings as particularly surprising: Rather than shedding light on the long-running debate over exactly how slowly the graph diameter *grows* as a function of the number of nodes, it suggests a need to revisit standard models so as to produce graphs in which the effective diameter is capable of actually *shrinking* over time. We also note that, while densification and decreasing diameters are properties that are intuitively consistent with one another (and are both borne out in the datasets we study), they are qualitatively distinct in the sense that it is possible to construct examples of graphs evolving over time that exhibit one of these properties but not the other.

We can further sharpen the quantitative aspects of these findings. In particular, the densification of these graphs, as suggested by (A'), is not arbitrary; we find that as the graphs evolve over time, they follow a version of the relation

$$e(t) \propto n(t)^a \tag{1}$$

where $e(t)$ and $n(t)$ denote the number of edges and nodes of the graph at time t, and a is an exponent that generally lies strictly between 1 and 2. We refer to such a relation as a *densification power law*, or *growth power law*. (Exponent $a = 1$ corresponds to constant average degree over time, while $a = 2$ corresponds to an extremely dense graph where each node has, on average, edges to a constant fraction of all nodes.)

What underlying process causes a graph to systematically densify, with a fixed exponent as in Equation (1), and to experience a decrease in effective diameter even as its size increases? This question motivates the second main contribution of this work: we present two families of probabilistic generative models for graphs that capture aspects of these properties. The first model, which we refer to as *Community Guided Attachment* (CGA), argues that graph densification can have a simple underlying basis; it is based on a decomposition of the nodes into a nested set of communities, such that the difficulty of forming links between communities increases with the community size. For this model, we obtain rigorous results showing that a natural tunable parameter in the model can lead to a densification power law with any desired exponent a. The second model, which is more sophisticated, exhibits both densification and a decreasing effective diameter as it grows. This model, which we refer to as the *Forest Fire Model*, is based on having new nodes attach to the network by "burning" through existing edges in epidemic fashion. The mathematical analysis of this model appears to lead to novel questions about random graphs that are quite complex, but through simulation we find that for a range of parameter values the model exhibits realistic behavior in densification, distances, and degree distributions. It is thus the first model, to our knowledge, that exhibits this full set of desired properties.

Accurate properties of network growth, together with models supporting them, have implications in several contexts.

- *Graph generation:* Our findings form means for assessing the quality of graph generators. Synthetic graphs are important for 'what if' scenarios, for extrapolations, and for simulations, when real graphs are impossible to collect (like, e.g., a very large friendship graph between people).
- *Graph sampling:* Datasets consisting of huge real-world graphs are increasingly available, with sizes ranging from the millions to billions of nodes. There are many known algorithms to compute interesting measures (shortest paths, centrality, betweenness, etc), but most of these algorithms become impractical for large graphs. Thus sampling is essential — but sampling from a graph is a non-trivial problem. Densification laws can help discard bad sampling methods, by providing means to reject sampled subgraphs.
- *Extrapolations:* For several real graphs, we have a lot of snapshots of their past. What can we say about their future? Our results help form a basis for validating scenarios for graph evolution.
- *Abnormality detection and computer network management:* In many network settings, "normal" behavior will produce subgraphs that obey densification laws (with a predictable exponent) and other properties of network growth. If we detect activity producing structures that deviate significantly from this, we can flag it as an abnormality; this can potentially help with the detection of e.g. fraud, spam, or distributed denial of service (DDoS) attacks.

The rest of the paper is organized as follows: Section 2 surveys the related work. Section 3 gives our empirical findings on real-world networks across diverse domains. Section 4 describes our proposed models and gives results obtained both through analysis and simulation. We conclude and discuss the implications of our findings in Section 5.

2. RELATED WORK

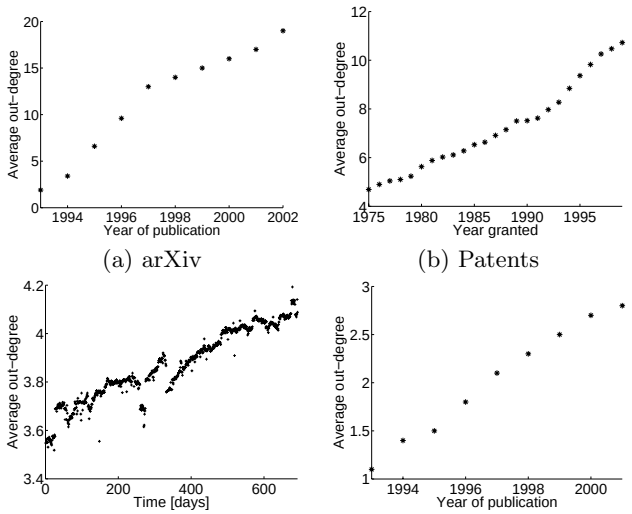
Research over the past few years has identified classes of properties that many real-world networks obey. One of the main areas of focus has been on *degree power laws*, showing that the set of node degrees has a heavy-tailed distribution. Such degree distributions have been identified in phone call 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*.

This data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP–TH section. For each month m ($1 \leq m \leq 124$) we create a citation graph using all papers published before month m. For each of these graphs, we plot the number of nodes versus the number of edges on a logarithmic scale and fit a line.

Figure 2(a) shows the DPL plot; the slope is $a = 1.68$ and corresponds to the exponent in the densification law. Notice that a is significantly higher than 1, indicating a large deviation from linear growth. As noted earlier, when a graph has $a > 1$, its average degree increases over time. Figure 1(a) exactly plots the average degree $\bar{d}$ over time, and it is clear that $\bar{d}$ increases. This means that the average length of the bibliographies of papers increases over time.

There is a subtle point here that we elaborate next: With almost any network dataset, one does not have data reaching all the way back to the network's birth (to the extent that this is a well-defined notion). We refer to this as the problem of the "*missing past*." Due to this, there will be some effect of increasing out-degree simply because edges will point to nodes prior to the beginning of the observation period. We refer to such nodes as *phantom nodes*, with a similar definition for *phantom edges*. In all our datasets, we find that this effect is relatively minor once we move away from the beginning of the observation period; on the other hand, the phenomenon of increasing degree continues through to the present. For example, in arXiv, nodes over the most recent years are primarily referencing non-phantom nodes; we observe a knee in Figure 1(a) in 1997 that appears to be attributable in large part to the effect of phantom nodes. (Later, when we consider a graph of the Internet, we will see a case where comparable properties hold in the absence of any "missing past" issues.)

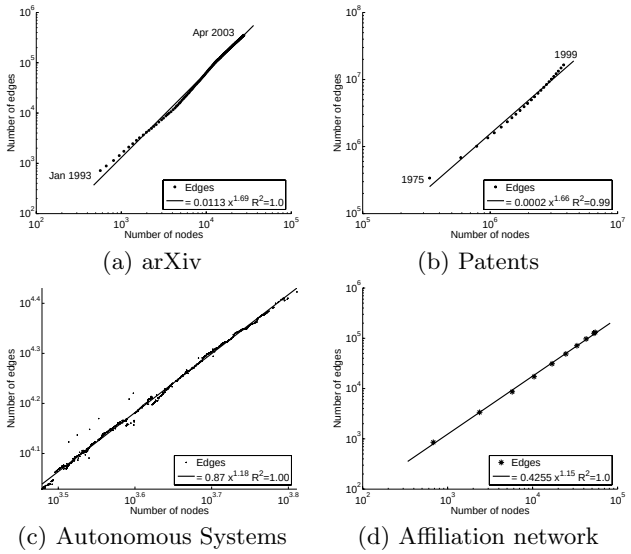
We also experimented with a second citation graph, taken from the HEP–PH section of the arXiv, which is about the same size as our first arXiv dataset. It exhibits the same behavior, with the densification exponent $a = 1.56$. The plot is omitted for brevity.

3.1.2 Patents citation graph

Next, we consider a U.S. patent dataset maintained by the National Bureau of Economic Research [13]. The data set spans 37 years (January 1, 1963 to December 30, 1999), and includes all the utility patents granted during that period, totaling n=3,923,922 patents. The citation graph includes all citations made by patents granted between 1975 and 1999, totaling e=16,522,438 citations. Because the dataset begins in 1975, it too has a "missing past" issue, but again the effect of this is minor as one moves away from the first few years.

We follow the same procedure as with arXiv. For each year Y from 1975 to 1999, we create a citation network on patents up to year Y, and give the DPL plot, in Figure 2(b). As with the arXiv citation network, we observe a high densification exponent, in this case $a = 1.66$.

Figure 1(b) illustrates the increasing out-degree of patents over time. Note that this plot does not incur any of the complications of a bounded observation period, since the patents in the dataset include complete citation lists, and here we are simply plotting the average size of these as a function of the year.



(a) arXiv (b) Patents

(c) Autonomous Systems (d) Affiliation network

Figure 2: Number of edges $e(t)$ versus number of nodes $n(t)$, in log-log scales, for several graphs. All 4 graphs obey the Densification Power Law, with a consistently good fit. Slopes: a = 1.68, 1.66, 1.18 and 1.15, respectively.

3.1.3 Autonomous systems graph

The graph of routers comprising the Internet can be organized into sub-graphs called Autonomous Systems (AS). Each AS exchanges traffic flows with some neighbors (peers). We can construct a communication network of who-talks-to-whom from the BGP (Border Gateway Protocol) logs.

We use the the *Autonomous Systems (AS)* dataset from [26]. The dataset contains 735 daily instances which span an interval of 785 days from November 8 1997 to January 2 2000.

In contrast to citation networks, where nodes and edges only get added (not deleted) over time, the AS dataset also exhibits both the addition and deletion of the nodes and edges over time.

Figure 2(c) shows the DPL plot for the Autonomous Systems dataset. We observe a clear trend: Even in the presence of noise, changing external conditions, and disruptions to the Internet we observe a strong super-linear growth in the number of edges over more than 700 AS graphs. We show the increase in the average node degree over time in Figure 1(c). The densification exponent is $a = 1.18$, lower than the one for the citation networks, but still clearly greater than 1.

3.1.4 Affiliation graphs

Using the arXiv data, we also constructed bipartite *affiliation graphs*. There is a node for each paper, a node for each person who authored at least one arXiv paper, and an edge connecting people to the papers they authored. Note that the more traditional *co-authorship network* is implicit in the affiliation network: two people are co-authors if there is at least one paper joined by an edge to each of them.

We studied affiliation networks derived from the five largest categories in the arXiv (ASTRO–PH, HEP–TH, HEP–PH, COND–MAT and GR–QC). We place a time-stamp on each node: the submission date of each paper, and for each per-

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 $a = 1.15$.

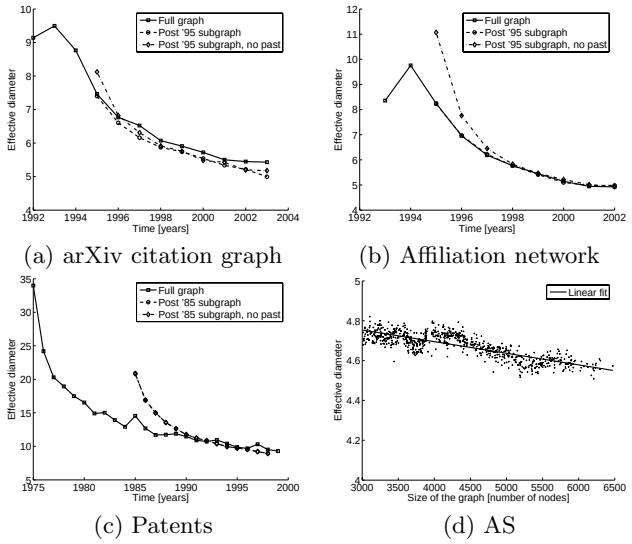
3.2 Shrinking Diameters

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

giant component; again, our results are essentially the same using these two methods.

"Missing Past" effects: A third issue is the problem of the "missing past," the same general issue that surfaced in the previous subsection when we considered densification. In particular, we must decide how to handle citations to papers that predate our earliest recorded time. (Note that the missing past is not an issue for the AS network data, where the effective diameter also decreases.)

To understand how the diameters of our networks are affected by this unavoidable problem, we perform the following test. We pick some positive time $t_0 > 0$, and determine what the diameter would look like as a function of time, *if this were the beginning of our data*. We can then put back in the nodes and the edges from before time t_0, and study how much the diameters change. If this change is small — or at least if it does not affect the qualitative conclusions — then it provides evidence that the missing past is not influencing the overall result.

Specifically, we set this cut-off time t_0 to be the beginning of 1995 for the arXiv (since we have data from 1993), and to be 1985 for the patent citation graph (we have data from 1975). We then compared the results of three measurements:

• **Diameter of full graph.** We compute the effective diameter of the graph's giant component.

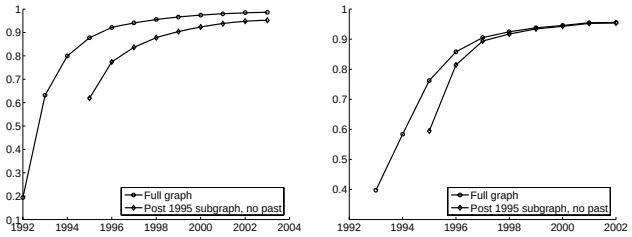
• **Post-t_0 subgraph.** We compute the effective diameter of the post-t_0 subgraph using all nodes and edges. This means that for each time t ($t > t_0$) we create a graph using all nodes dated before t. We then compute the effective diameter of the subgraph of the nodes dated between t_0 and t. To compute the effective diameter we can use all edges and nodes (including those dated before t_0). This experiment measures the diameter of the graph if we knew the full (pre-t_0) past — the citations of the papers which we have intentionally excluded for this test.

• **Post-t_0 subgraph, no past.** We set t_0 the same way as in previous experiment, but then for all nodes dated before t_0 we delete all their out-links. This creates the graph we would have gotten if we had started collecting data only at time t_0.

In Figure 3, we superimpose the effective diameters using the three different techniques. If the missing past does not play a large role in the diameter, then all three curves should lie close to one another. We observe this is the case for the arXiv citation graphs. For the arXiv paper-author affiliation graph, and for the patent citation graph, the curves are quite different right at the cut-off time t_0 (where the effect of deleted edges is most pronounced), but they quickly align with one another. As a result, it seems clear that the continued decreasing trend in the effective diameter as time runs to the present is not the result of these boundary effects.

Emergence of giant component: A final issue is the dynamics by which the giant component emerges. For example, in the standard Erdös-Renyi random graph model (which has a substantially different flavor from the growth dynamics of the graphs here), the diameter of the giant component is quite large when it first appears, and then it shrinks as edges continue to be added. Could shrinking diameters in some way be a symptom of emergence of giant component?

It appears fairly clear that this is not the case. Figure 4 shows the fraction of all nodes that are part of the largest connected component (GCC) over time. We plot the size of the GCC for the full graph and for a graph where we had no



(a) arXiv citation graph (b) Affiliation network

Figure 4: The fraction of nodes that are part of the giant connected component over time. We see that after 4 years the 90% of all nodes in the graph belong to giant component.

past — i.e., where we delete all out-links of the nodes dated before the cut-off time t_0. Because we delete the out-links of the pre-t_0 nodes the size of GCC is smaller, but as the graph grows the effect of these deleted links becomes negligible.

We see that within a few years the giant component accounts for almost all the nodes in the graph. The effective diameter, however, continues to steadily decrease beyond this point. This indicates that the decrease is happening in a "mature" graph, and not because many small disconnected components are being rapidly glued together.

Based on all this, we believe that the decreasing diameters in our study reflect a fundamental property of the underlying networks. Understanding the possible causes of this property, as well as the causes of the densification power laws discussed earlier, will be the subject of the next section.

4. PROPOSED MODELS

We have now seen that densification power laws and shrinking effective diameters are properties that hold across a range of diverse networks. Moreover, existing models do not capture these phenomena. We would like to find some simple, local model of behavior, which could naturally lead to the macroscopic phenomena we have observed. We present increasingly sophisticated models, all of which naturally achieve the observed densification; the last one (the "Forest Fire" model) also exhibits shrinking diameter and all the other main patterns known (including heavy-tailed in- and out-degree distributions).

4.1 Community Guided Attachment

What are the underlying principles that drive all our observed graphs to obey a densification power law, without central control or coordination? We seek a model in which the densification exponent arises from intrinsic features of the process that generates nodes and edges. While one could clearly define a graph model in which $e(t) \propto n(t)^a$ by simply having each node, when it arrives at time t, generate $n(t)^{a-1}$ out-links — the equivalent of positing that each author of a paper in a citation network has a rule like, "Cite n^{a-1} other documents," hard-wired in his or her brain — such a model would not provide any insight into the origin of the exponent a, as the exponent is unrelated to the operational details by which the network is being constructed. Instead, our goal is to see how underlying properties of the network evolution process itself can affect the observed densification behavior.

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:*

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

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 $(1, 2]$.

COROLLARY 1. *If the Difficulty Function is scale-free ($f(h) = c^{-h}$, with $1 < c < b$), then the Community Guided Attachment obeys the Densification Power Law with exponent*

$$a = 2 - \log_b(c)$$

4.1.2 Dynamic Community Guided Attachment

So far we have discussed a model in which nodes are first organized into a nested set of communities, and then they start forming links. We now extend this to a setting in which nodes are added over time, and the nested structure deepens to accommodate them. We will assume that a node only creates out-links at the moment it is added (and hence, only to nodes already present); this is natural for domains like citation networks in which a paper's citations are written at the same time as the paper itself.

Specifically, the model is as follows. Rather than having graph nodes reside only at the leaves of the tree Γ, there will now be a graph node corresponding to every internal node of Γ as well. Initially, there is a single node v in the graph, and our tree Γ consists just of v. In time step t, we go from a complete b-ary tree of depth $t-1$ to one of depth t, by adding b new leaves as children of each current leaf. Each of these new leaves will contain a new node of the graph.

Now, each new node forms out-links according to a variant of the process in Section 4.1.1. However, since a new node has the ability to link to internal nodes of the existing tree, not just to other leaves, we need to extend the model to incorporate this. Thus, we define the *tree-distance* $d(v, w)$ between nodes v and w to be the length of a path between them in Γ — this is the length of the path from v up to the least common ancestor of v and w, plus the length of the path from this least common ancestor down to w. Note that if v and w are both leaves, then $d(v, w) = 2h(v, w)$, following the definition of $h(v, w)$ given previously.

The process of forming out-links is now as follows: For a constant c, node v forms a link to each node w, independently, with probability $c^{-d(v,w)/2}$. (Note that dividing by 2 in the exponent means this model gives the same probability as basic model in the case when both v and w are leaves.)

Like the first model, this process produces a densification law with exponent $a = 2 - \log_b(c)$ when $c < b$. However, for $c < b^2$, it also yields a heavy-tailed distribution of in-degrees — something that the basic model did not produce. We describe this in the following theorem; due to space limitations, we omit the proof from this version of the paper.

THEOREM 2. *The dynamic Community Guided Attachment model just defined has the following properties.*

- *When $c < b$, the average node degree is $n^{1-\log_b(c)}$ and the in-degrees follow a Zipf distribution with exponent $\frac{1}{2}\log_b(c)$.*
- *When $b < c < b^2$, the average node degree is constant, and the in-degrees follow a Zipf distribution with exponent $1 - \frac{1}{2}\log_b(c)$.*
- *When $c > b^2$, the average node degree is constant and the probability of an in-degree exceeding any constant bound k decreases exponentially in k.*

Thus, the dynamic Community Guided Attachment model exhibits three qualitatively different behaviors as the parameter c varies: densification with heavy-tailed in-degrees; then constant average degree with heavy-tailed in-degrees; and then constant in- and out-degrees with high probability. Note also the interesting fact that the Zipf exponent is maximized for the value of c right at the onset of densification.

Finally, we have experimented with versions of the dynamic Community Guided Attachment model in which the tree is not balanced, but rather deepens more on the left branches than the right (in a recursive fashion). We have also considered versions in which a single graph node can "reside" at two different nodes of the tree Γ, allowing for graph nodes to be members of different communities. We do not go into further details of these extensions in this version of the paper.

4.2 The Forest Fire Model

Community Guided Attachment and its extensions show how densification can arise naturally, and even in conjunction with heavy-tailed in-degree distributions. However, it is not a rich enough class of models to capture all the properties in our network datasets. In particular, we would like to capture both the shrinking effective diameters that we have observed, as well as the fact that real networks tend to have heavy-tailed out-degree distributions (though generally not as skewed as their in-degree distributions). The Community Guided Attachment models do not exhibit either of these properties.

Specifically, our goal is as follows. Given a (possibly empty) initial graph G, and a sequence of new nodes $v_1 \ldots v_k$, we want to design a simple randomized process to successively link v_i to nodes of G ($i = 1, \ldots k$) so that the resulting graph G_{final} will obey all of the following patterns: heavy-tailed distributions for in- and out-degrees, the Densification Power Law, and shrinking effective diameter.

We are guided by the intuition that such a graph generator may arise from a combination of the following components:

- some type of "rich get richer" attachment process, to lead to heavy-tailed in-degrees;
- some flavor of the "copying" model [19], to lead to communities;
- some flavor of Community Guided Attachment, to produce a version of the Densification Power Law;
- and a yet-unknown ingredient, to lead to shrinking diameters.

Note that we will not be assuming a community hierarchy on nodes, and so it is not enough to simply vary the Community Guided Attachment model.

Based on this, we introduce the *Forest Fire Model*, which is capable of producing all these properties. To set up this model, we begin with some intuition that also underpinned Community Guided Attachment: nodes arrive in over time; each node has a "center of gravity" in some part of the network; and its probability of linking to other nodes decreases rapidly with their distance from this center of gravity. However, we add to this picture the notion that, occasionally, a new node will produce a very large number of out-links. Such nodes will help cause a more skewed out-degree distribution; they will also serve as "bridges" that connect formerly disparate parts of the network, bringing the diameter down.

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 two random numbers: x and y that are geometrically distributed with means $(1-p)^{-1}$ and $(1-rp)^{-1}$ respectively. Node v selects x out-links and y in-links incident to nodes that were not yet visited. Let $w_1, w_2, \ldots, w_{x+y}$ denote the other ends of these selected links. If not enough in- or out-links are available, v selects as many is it can.

(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.

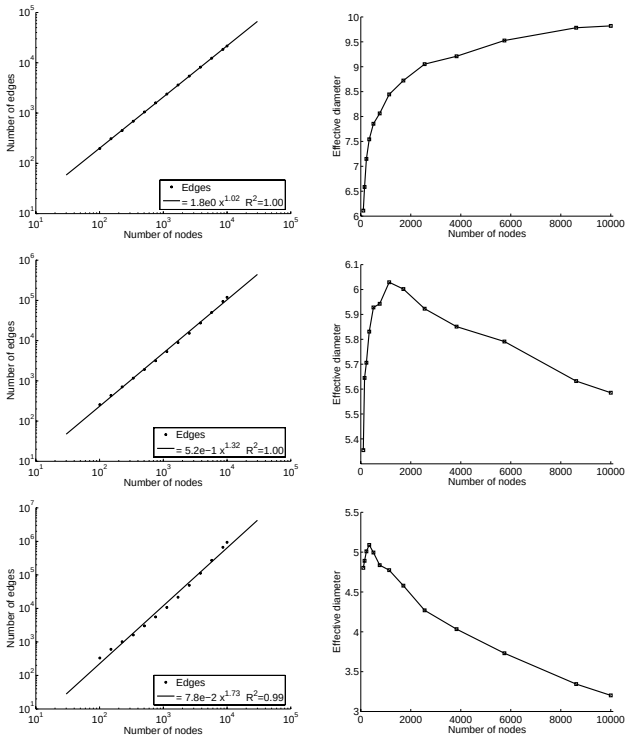
Rigorous analysis of the Forest Fire Model appears to be quite difficult. However, in simulations, we find that by varying just the two parameters p and r, we can produce graphs that densify ($a > 1$), exhibit heavy-tailed distributions for both in- and out-degrees (Fig. 6), and have diameters that decrease. 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 Patents 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 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



DPL exponent Diameter

Figure 5: The DPL plot and diameter for Forest Fire model. Top: sparse graph ($a = 1.01 < 2$), with increasing diameter (forward burning probability: $p = 0.35$, backward probability: $p_b = 0.20$). Middle: (most realistic case:) densifying graph ($a = 1.32 < 2$) with decreasing diameter ($p = 0.37$, $p_b = 0.33$). Bottom: dense graph ($a \approx 2$), with decreasing diameter ($p = 0.38$, $p_b = 0.35$).

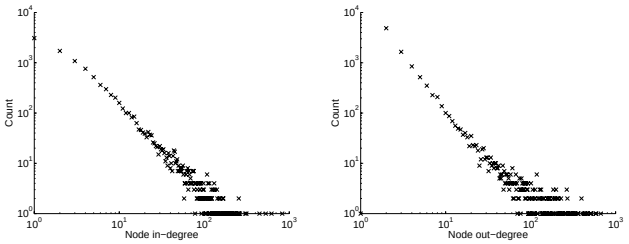
large distances caused by groups of nodes linking to different orphans gradually diminishing as further nodes arrive to connect them together.

Multiple ambassadors: We experimented with allowing newcomers to choose more than one ambassador with some positive probability. That is, rather than burning links starting from just one node, there is some probability that a newly arriving node burns links starting from two or more. This extension also accentuates the decrease in effective diameter over time, as nodes linking to multiple ambassadors serve to bring together formerly far-apart parts of the graph.

4.2.3 Phase plot

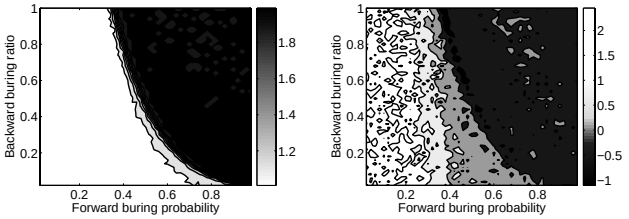
In order to understand the densification and diameter properties of graphs produced by the Forest Fire Model, we have explored the full parameter space of the basic model in terms of its two underlying quantities: the forward burning probability p and the backward burning ratio r.

Figure 7 shows how the densification exponent and the effective diameter depend on the values of these parameters. The densification exponent a is computed as in Section 3, by fitting a relation of the form $e(t) \propto n(t)^a$. For the effective diameter, we fit a logarithmic function of the form $diameter = \alpha \log t + \beta$ (where t is the current time, and



In-degree Out-degree

Figure 6: Degree distribution of a sparse graph with decreasing diameter (forward burning probability: 0.37, backward probability: 0.32).



Densification exponent Diameter

Figure 7: Contour plots: The Densification Power Law exponent a (left) and diameter log-fit α factor (right) over the parameter space (forward-burning probability and ratio).

hence the current number of vertices) to the last half of the effective diameter plot; we then report the coefficient α. Thus, $\alpha < 0$ corresponds to decreasing effective diameter over time.

Figure 7(a) gives the contour plot of the densification exponent a. The white color is for $a = 1$ (the graph maintains constant average degree), while the black color is for $a = 2$ (the graph is "dense", that is, the number of edges grows quadratically with the number of nodes, as, e.g., in the case of a clique). The desirable grey region is in-between; we observe that it is very narrow: a increases dramatically along a contour line, suggesting a sharp transition.

Figure 7(b) gives the contour plot for the factor α in the effective diameter fit, as defined above. The boundary between decreasing and increasing effective diameter is shifted somewhat from the contour line for densification, indicating that even the basic Forest Fire Model is able to produce sparse graphs (with densification exponent near 1) in which the effective diameter decreases.

For lack of space, we omit the phase plots with orphans and multiple ambassadors, which show similar behavior.

5. CONCLUSION

Despite the enormous recent interest in large-scale network data, and the range of interesting patterns identified for static snapshots of graphs (e.g. heavy-tailed distributions, small-world phenomena), there has been relatively little work on the properties of the time evolution of real graphs. This is exactly the focus of this work. The main findings and contributions follow:

• 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.