Affiliation Networks

Silvio Lattanzi*
Sapienza University
Rome, Italy
lattanzi@di.uniroma1.it

D. Sivakumar
Google Inc.
Mountain View, USA
siva@google.com

ABSTRACT

In the last decade, structural properties of several naturally arising networks (the Internet, social networks, the web graph, etc.) have been studied intensively with a view to understanding their evolution. In recent empirical work, Leskovec, Kleinberg, and Faloutsos identify two new and surprising properties of the evolution of many real-world networks: densification (the ratio of edges to vertices grows over time), and shrinking diameter (the diameter reduces over time to a constant). These properties run counter to conventional wisdom, and are certainly inconsistent with graph models prior to their work.

In this paper, we present the first model that provides a simple, realistic, and mathematically tractable generative model that intrinsically explains all the well-known properties of the social networks, as well as densification and shrinking diameter. Our model is based on ideas studied empirically in the social sciences, primarily on the groundbreaking work of Breiger (1973) on bipartite models of social networks that capture the affiliation of agents to societies.

We also present algorithms that harness the structural consequences of our model. Specifically, we show how to overcome the bottleneck of densification in computing shortest paths between vertices by producing sparse subgraphs that preserve or approximate shortest distances to all or a distinguished subset of vertices. This is a rare example of an algorithmic benefit derived from a realistic graph model.

Finally, our work also presents a modular approach to connecting random graph paradigms (preferential attachment, edge-copying, etc.) to structural consequences (heavy-tailed degree distributions, shrinking diameter, etc.).

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics. Stochastic processes.

*Research done while visiting Google, Inc., Mountain View. This work is partially supported by the Italian Ministry of University and Research under Cofin 2006 grant WebRam.

STOC'09, May 31–June 2, 2009, Bethesda, Maryland, USA.

General Terms

Algorithms, Theory.

Keywords

Social Networks, densification, diameter, sparsification.

1. INTRODUCTION

In the past few years, the idea of *networks* as a unifying theme to study "how social, technological, and natural worlds are connected" has emerged as an important and extroverted direction within theoretical CS. For a delightful example that contains a sample of the topics, see the web page for the Cornell course entitled "Networks", developed by Easley and Kleinberg.

Within this umbrella, the specific aim of this paper is to develop mathematical models of real-world "social" networks that are *realistic*, *mathematically tractable*, and — perhaps most importantly — *algorithmically useful*. There are several models of social networks that are natural and realistic (fit available data) but are hard from an analytical viewpoint; the ones that are amenable to mathematical analysis or that have algorithmic significance are often unnatural or unrealistic. In contrast, we present a model, rooted in sociology, that leads to clean mathematical analysis as well as algorithmic benefits.

We now briefly outline the history of significant recent developments in modeling real-world networks that provide the immediate context for our work. The numerous references from and to these salient pieces of work will offer the reader a more comprehensive picture of this area.

Internet and Web Graphs. One of the first observations that led to the interest in random graph models significantly different from the classical Erdős–Rényi models comes in the work of Faloutsos et al.[16], who noticed that the degree distribution of the Internet graph[1] is heavy-tailed, and roughly obeys a "power law," that is, for some constant $\alpha > 0$, the fraction of nodes of degree d is proportional to $d^{-\alpha}$. Similar observations were made about the web graph[2] by Barabasi and Albert [2], who also presented models based on the notion of "preferential attachment," wherein a network evolves by new nodes attaching themselves to existing nodes with probability proportional to the degrees of those

[1]Loosely speaking, this is the graph whose vertices are computers and whose edges are network links.
[2]This is the graph whose vertices are web pages, and whose directed edges are hyperlinks among web pages.

nodes. Both works draw their inspiration and mathematical precedents from classical works of Zipf [33], Mandelbrot [24], and Simon [30]. The latter work was formalized and studied rigorously in [6, 12, 7]. Broder et al. [10] made a rich set of observations about the degree and connectivity structure of the web graph, and showed that besides power-law degree distribution, the web graph consisted of numerous dense bipartite subgraphs (often dubbed "communities").

Within theoretical CS, Aiello et al. [1], and Kumar et al. [19] presented two models of random graphs, both of which offer rigorous explanations for power-law degree distributions; the models of [19] also led to graphs with numerous dense bipartite subgraphs, the first models to do so. The models of [19] are based on the notion of graph evolution by "copying," where each new vertex picks an existing vertex as its "prototype," and copies (according to some probabilistic model) its edges.

Preferential attachment and edge copying are two basic paradigms that both lead to heavy-tailed degree distributions and small diameter. The former is simpler to analyze, and indeed, despite its shortcomings with respect to explaining community structure, it has been analyzed extensively [6, 7]. For an entirely different treatment, see [15].

Small-World Graphs. In another development, Watts and Strogatz [32], Kleinberg [18, 17], and Duncan and Watts [13] revisited a classic 1960's experiment of the sociologist Stanley Milgram [31], who discovered that, on average, pairs of people chosen at random from the population are only six steps apart in the network of first-name acquaintances. In Kleinberg's model, vertices reside in some metric space, and a vertex is usually connected to most other vertices in its metric neighborhood, and, in addition, to a few "long range" neighbors. Kleinberg introduced an algorithmic twist, and proved the remarkable result that the network has small diameter and easily discoverable paths iff the long-range neighbors are chosen in a specific way.

Kleinberg's models, dubbed "small-world networks," offer a nice starting point to analyze social networks[3]. A piece of folklore wisdom about social networks is the observation that friendship is mostly transitive, that is, if a and b are friends and b and c are friends, then there is a good chance that a and c are friends as well. Kleinberg's model certainly produces graphs that satisfy this condition, but because of its stylized nature, isn't applicable in developing an understanding of real social networks. The other limitation of Kleinberg's model is that it is static, and is not a model of graph evolution.

Densification and Shrinking Diameter. Returning to the topic of evolving random graphs, the next significant milestone is the work of Leskovec et al.[22], who made two stunning empirical observations, both of which immediately invalidate prior models based on preferential attachment, edge copying, etc., as well as the small-world models. Namely, they reported that real-world networks became *denser* over time (super-constant average degree), and their diameters effectively *decreased* over time!

[3]Collaboration networks among authors, email and instant messaging networks, as well as the ones underlying Friendster, LiveJournal, Orkut, LinkedIn, MySpace, FaceBook, Bebo, etc. Indeed, the work of [28] demonstrates interesting correlations of friendships on the LiveJournal network with *geographic* proximity as an underlying metric for a small-world model.

The dual pursuits of empirical observations and theoretical models go hand in hand[4], and the work of [22] poses new challenges for mathematical modeling of real-world networks. Along with their observations, Leskovec et al. [22] present two graph models called the "community guided attachment" and "forest fire model". The former is a hierarchical model, and the latter is based on an extension of edge copying. While several analytical results are proved concerning these two models in [22], the models are quite complex and do not admit analyses powerful enough to establish all the observed properties, most notably degree distribution, densification, and shrinking diameter simultaneously.

The papers [21] and [23] study models explicitly contrived to be mathematically tractable and yielding the observed properties, without any claims of being realistic or intuitively natural. In the opposite direction, Leskovec et al. [20] propose a model that fit the data quite well, but that do not admit mathematical analyses. The crucial features of the latter model are that edges are created based on preferential attachment and by randomly "closing triangles."

Affiliation Networks. Our design goals for a mathematical model of generic social networks are that it should be simple to state and intuitively natural, sufficiently flexible and modular in structure with respect to the paradigms employed, and, of course, by judicious choice of the paradigms, offer compelling explanations of the empirically observed phenomena.

The underlying idea behind our model is that in social networks there are two types of entities — *actors* and *societies* — that are related by affiliation of the former in the latter. These relationships can be naturally viewed as bipartite graphs, called *affiliation networks*; the social network among the actors that results from the bipartite graph is obtained by "folding" the graph, that is, replacing paths of length two in the bipartite graph among actors by an (undirected) edge. The central thesis in developing a social network as an folded affiliation network is that acquaintanceships among people often stem from one or more common or shared affiliations — living on the same street, working at the same place, being fans of the same football club, having coauthored a paper together, etc.

Affiliation networks are certainly not new — indeed, this terminology is prevalent in sociology, and a fundamental 1974 paper of Breiger [9] appears to be the first one to explicitly address the duality of "persons and groups" in the context of "networks of interpersonal ties... [and] intergroup ties." Breiger notes that the metaphor of this "dualism" occurs as early as in 1902 in the work of Cooley.

Our model for the evolving affiliation network and the consequent social network incorporates elements of preferential attachment and edge copying in fairly natural ways. The folding rule we analyze primarily in the paper is the one that places an undirected edge between every pair of (actor) nodes connected by a length-2 path in the bipartite graph. We consider some extensions for which our analyses continue to work, and more generally, on the flexibility of our model in Section 8. We show that when an affiliation network B is generated according to our model and its folding G on n vertices is produced, the resulting graphs satisfy the following properties:

[4]See Mitzenmacher's editorial [27] for an eloquent articulation of this phenomenon

(1) B has a power-law distribution, and G has a heavy-tailed degree distribution as well, and all but $o(n)$ vertices of G have bounded degree;

(2) under a mild condition on the ratio of the expected degree of actor nodes and society nodes in B, the graph G has superlinear number of edges;

(3) under the same condition, the effective diameter of G stabilizes to a constant.

Algorithmic Benefits, and an Application. Although they are very interesting, these structural properties do not yield any direct insight into the development of efficient algorithms for challenging problems on large-scale graphs. With our model of networks based on affiliation graphs, we take a significant step towards remedying this situation. We show how we can approach path problems on our networks by taking advantage of a key feature in their structure. Namely, we utilize the fact that even though the ultimate network produced by the model is dense, there is a sparse (constant average degree) backbone of the network given by the underlying affiliation network.

First we show that if we are given a large random set R of distinguished nodes and we care about paths from arbitrary nodes to nodes in R, then we can sparsify the graph to have only a small constant fraction of its edges, yet preserving all shortest distances to vertices in R. Secondly, we show that if we are allowed some distortion, we can sparsify the graph significantly via a simple algorithm for graph spanners: namely, we show that we can sparsify the graph to have a *linear* number of edges, while stretching distances by no more than a factor given by the ratio of the expected degree of actor and society nodes in the affiliation network.

Finally, we mention our motivating example: a "social" network that emerges from search engine queries, where these shortest path problems have considerable significance. The affiliation network here is the bipartite graph of queries and web pages (urls), with edges between queries and urls that users clicked on for the query; by folding this network, we may produce a graph on just the queries, whose edges take on a natural semantics of relatedness. Now suppose we are given a distinguished subset of queries that possess some significance (high commercial value, topic names, people names, etc.). Given any query, we can: find the nearest commercial queries (to generate advertisements), classify the query into topics, or discover people associated with the query. We have empirically observed that our sparsification algorithms work well on these graphs with hundreds of millions of nodes.

Critique of our work. In this paper we only analyze the most basic folding rule, namely replace each society node in the affiliation network by a complete graph on its actors in the folded graph. As noted in Section 8, this could be remedied somewhat without losing the structural properties; we leave for future work a more detailed exploration of the possibilities here.

The next drawback of our models is that given a social network (or other large graph), it is not at all clear how one can test the hypothesis that it was formed by the folding of an affiliation network. The general problem of solving, given a graph G on a set Q of vertices, whether it was obtained by folding an affiliation network on vertex sets Q and U, where $|U| = O(|Q|)$, is NP-Complete.

Finally, our model of folded affiliation networks seems limited to social networks among people related together by various attributes (the societies). A feature that is often seen in several large real networks that appears to be missed by our model is the presence of an approximately hierarchical structure (for example, the Internet graph exhibits an approximate hierarchy in the form of autonomous systems, domains, intra- and inter-domain edges via gateways, and so forth).

2. OUR MODEL

In our model, two graphs evolve at the same time. The first one is a simple bipartite graph that represents the affiliation network; we refer to this graph as $B(Q, U)$. The second one it is the social network graph, we call this graph $G(Q, E)$. The set Q is the same in both graphs. As defined, $G(Q, E)$ is a multigraph, so we also analyze the underlying simple graph $\hat{G}(Q, \hat{E})$; the results and proofs for $\hat{G}(Q, \hat{E})$ are omitted. For readability, we present the two evolution processes separately even though the two graphs evolve together.

$B(Q, U)$
Fix two integers $c_q, c_u > 0$, and let $\beta \in (0, 1)$.
At time 0, the bipartite graph $B_0(Q, U)$ is a simple graph with at least $c_q c_u$ edges, where each node in Q has at least c_q edges and each node in U has at least c_u edges.
At time $t > 0$:
(Evolution of Q) With probability β:
(*Arrival*) A new node q is added to Q.
(*Preferentially chosen Prototype*) A node $q' \in Q$ is chosen as *prototype* for the new node, with probability proportional to its degree.
(*Edge copying*) c_q edges are "copied" from q'; that is, c_q neighbors of q', denoted by $u_1, \ldots, u_{c_q}$, are chosen uniformly at random (without replacement), and the edges $(q, u_1), \cdots, (q, u_{c_q})$ are added to the graph.
(Evolution of U) With probability $1 - \beta$, a new node u is added to U following a symmetrical process, adding c_u edges to u.

In order to understand the intuition behind this evolving process, let us consider, for example, the citation graph among papers. In this case the bipartite graph consists of papers, the set Q, and topics, the set U. Now when an author writes a new paper, he probably has in mind some older paper that will be the *prototype*, and he is likely to write on (a subset of the) topics considered in this *prototype*. Similarly, when a new topic emerges in the literature, it is usually inspired by an existing topic (prototype) and it has been probably foreseen by older papers.

We call *folded* any edge that is in $G_0(Q, E)$ or has been added to $G(Q, E)$ via the prototype or by evolution of U; the set of folded edges is denoted by F.

To continue the example of citation networks, the intuition behind the construction of $G(Q, E)$ is that when an author writes the references of a new paper he will cite all, or most, of the paper on the same topics and some other papers of general interest. The same ideas that suggest this model for the citation graph can be applied also to other social graphs.

$G(Q,E)$
Fix integers $c_q, c_u, s > 0$, and let $\beta \in (0,1)$.
At time 0, $G_0(Q,E)$ consists of the subset Q of the vertices of $B_0(Q,U)$, and two vertices have an edge between them for every neighbor in U that they have in common in $B_0(Q,U)$.
At time $t > 0$:
(Evolution of Q) With probability β:
(*Arrival*) A new node q is added to Q.
(*Edges via Prototype*) An edge between q and another node in Q is added for every neighbor that they have in common in $B(Q,U)$ (note that this is done after the edges for q are determined in B).
(Edges via evolution of U) With probability $1-\beta$:
A new edge is added between two nodes q_1 and q_2 if the new node added to $u \in U$ is a neighbor of both q_1 and q_2 in $B(Q,U)$.
(Preferentially Chosen Edges) A set of s nodes $q_{i_1}, \ldots, q_{i_s}$ is chosen, each node independently of the others (with replacement), by choosing vertices with probability proportional to their degrees, and the edges $(q, q_{i_1}), \ldots, (q, q_{i_s})$ are added to $G(Q,E)$.

3. EVOLUTION OF THE DEGREE DISTRIBUTION OF AFFILIATION NETWORKS

In this section we state the main results concerning the degree distribution of $B(Q,U)$. The first result shows that the degree distribution of vertices in both Q and U satisfy power laws. In addition, we establish three additional lemmata that are important from a structural viewpoint; these lemmata tell us how the degrees of the high-degree vertices evolve, and also assert that most of the edges of B added "later" in the process have their end points pointing to a low-degree node.

THEOREM 1. *For the bipartite graph $B(Q,U)$ generated after n steps, almost surely, when $n \to \infty$, the degree sequence of nodes in Q (resp. U) follows a power law distribution with exponent $\alpha = -2 - \frac{c_q\beta}{c_u(1-\beta)}$ $\left(\alpha = -2 - \frac{c_u(1-\beta)}{c_q\beta}\right)$, for every degree smaller than n^γ, with $\gamma < \frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}$ $\left(\gamma < \frac{1}{4+\frac{c_u(1-\beta)}{c_q\beta}}\right)$.*

LEMMA 2. *If a node in $B(Q,U)$ has degree $g(n)$ at time n, with $g(n) \in \omega(\log n)$, it had degree with high probability $\Omega(g(n))$ also at time ϵn, for any constant $\epsilon > 0$.*

LEMMA 3. *If a node in $B(Q,U)$ has degree $\Theta(n^\lambda)$ at the end of the process a constant fraction of the node pointing to u have been inserted after time ϕn, for any constant $0 < \phi, \lambda < 1$.*

LEMMA 4. *At any time ϕn, for any constant $0 < \phi \leq 1$, the number of edges, in $B(Q,U)$, that points to a node in U of degree at least i is $\Theta\left(ni^{-\frac{c_u(1-\beta)}{c_q\beta}}\right)$, for any i up to n^γ, with $\gamma < \frac{1}{4+\frac{c_u(1-\beta)}{c_q\beta}}$.*

4. PROPERTIES OF THE DEGREE DISTRIBUTIONS OF THE GRAPHS $G(Q,E)$ AND $\hat{G}(Q,\hat{E})$

Although derived from $B(Q,U)$, the problem of computing the degree distributions of $G(Q,E)$ and of $\hat{G}(Q,\hat{E})$ is much harder; in this section we will show some interesting properties of the degree distribution of the folded graphs. First we will show that the probability of a random node G having high-degree dominates the complementary cumulative distribution function of the degree distribution of the nodes in U in $B(Q,U)$. Then, by construction, a similar theorem follows with respect to the nodes in Q. Together, these results imply:

THEOREM 5. *The degree distributions of the graphs $G(Q,E)$ and $\hat{G}(Q,\hat{E})$ are heavy-tailed.*

PROPOSITION 6. *For the folded graphs $G(Q,E)$ and $\hat{G}(Q,\hat{E})$ generated after n steps, almost surely, when $n \to \infty$, the complementary cumulative distribution function of the degrees of nodes inserted after time ϕn, for any constant $0 < \phi < 1$, dominates the complementary cumulative distribution of a power law with exponent $\alpha = -2 - \frac{c_u(1-\beta)}{c_q\beta}$, for every degree bigger than $\log^{2+\epsilon} n$ and smaller than n^γ, with $\gamma < \frac{1}{4+\frac{c_u(1-\beta)}{c_q\beta}}$.*

PROOF. Let Q^i be the number of nodes inserted after time ϕn and with degree at least i in $G(Q,E)$. Instead of computing directly Q^i we show that Q^i is bigger than a random variables, which is in $\Theta\left(ni^{-2-\frac{c_u(1-\beta)}{c_q\beta}}\right)$.

Let S^i be the number of edges inserted after time ϕn, pointing to a node of degree at least i in $B(Q,U)$, and such that $\forall (a,b) \in S^i$, if $a \in Q$ then (a,b) is the *oldest* edge pointing to the node a. By definition the following inequality holds: $S^i \leq Q^i$. Now by Lemma 2 we know that all the nodes inserted after time ϕn will have degree in $O(\log n)$ in $B(Q,U)$. So any node in Q^i has degree in $O(\log n)$ in $B(Q,U)$ and only c_u of its neighbors can have degree in $\omega(\log n)$. Hence if a node in Q^i has degree $i \in \omega(\log^{2+\epsilon} n)$ in $G(Q,E)$ at least one of its initial neighbors has degree in $\Omega(i)$ in $B(Q,U)$.

Now by Lemma 4 there are $\Theta\left(ni^{-\frac{c_u(1-\beta)}{c_q\beta}}\right)$ edges of degree at least i, for i up to n^γ, with $\gamma < \frac{1}{4+\frac{c_u(1-\beta)}{c_q\beta}}$, thus there are two constants p^* and $p_* \in \Theta\left(i^{-\frac{c_u(1-\beta)}{c_q\beta}}\right)$ such that $p_* < \Pr[\text{copying an edge of degree at least } i \text{ at time } t] < p^*$, for any $t \geq \phi n$.

Hence S^i dominates the number of heads that we have if we flip $\Theta((1-\phi)n)$ times a biased coin that gives head with probability p_*. Thus applying the Chernoff bound, we have: $\Theta(p_*(1-\phi)n) \leq Q^i$.

Hence $Q^i \in \Omega\left(ni^{-\frac{c_u(1-\beta)}{c_q\beta}}\right)$, and $\Pr[\text{a node in } Q' \text{ has degree} > i] \in \Omega\left(i^{-\frac{c_u(1-\beta)}{c_q\beta}}\right)$. □

PROPOSITION 7. *For the folded graphs $G(Q,E)$ and $\hat{G}(Q,\hat{E})$ generated after n steps, almost surely, when $n \to \infty$,*

the complementary cumulative distribution function of nodes dominates the complementary cumulative distribution function of a power law distribution with exponent $\alpha = -2 - \frac{c_q\beta}{c_u(1-\beta)}$*, for every degree smaller than* n^γ*, with* $\gamma < \frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}$.

Finally we will show that most the nodes have degrees in $\Theta(1)$.

PROPOSITION 8. *For the folded graphs* $G(Q,E)$ *and* $\hat{G}(Q,\hat{E})$ *generated after* n *steps, all but* $o(n)$ *nodes have degree in* $\Theta(1)$.

PROOF. We start by notice that we can restrict our attention to the edges in F because the edges in $|E-F| \in \Theta(n)$. Thus only $o(n)$ nodes can have degree in $\omega(1)$ in the graph $G(Q, E-F)$.
Further, by Theorem 1, all but $o(n)$ nodes in $B(Q,U)$ have degree $\in \Theta(1)$. In addition, recalling that for Lemma 4 only an $o(n)$ of the edges in $B(Q,U)$ will point to a non-constant degree node in U. We have that only an $o(n)$ of the nodes increase their degrees by more than a constant factor. □

5. DENSIFICATION OF EDGES

In this section we prove that the number of edges in the graph $G(Q,E)$ is $\omega(|Q|)$.

THEOREM 9. *If* $c_u < \frac{\beta}{1-\beta}c_q$ *the number of edges in* $G(Q,E)$ *is* $\omega(n)$.

PROOF. We notice that every node u in $U \in B(Q,U)$ in $G(Q,E)$ gives rise to a clique where all neighbors of u are connected. Thus we can lower bound the number of edges in the graph $G(Q,E)$ as follows:

$$\begin{aligned} |E| > & \sum_{i=1}^{n}(\#\text{ of nodes of degree } i \text{ in } U)\binom{i}{2} \\ \geq & \sum_{i=1}^{N}(\#\text{ of nodes of degree } i)\binom{i}{2}, \end{aligned}$$

where $N = n^\gamma$, with $\gamma < \frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}$. By Theorem 1 with high probability:

$$|E| > \sum_{i=1}^{N}\left(\left(\left(\frac{n}{\zeta\left(-2-\frac{c_u(1-\beta)}{c_q\beta}\right)}\frac{1}{i^{2+\frac{c_u(1-\beta)}{c_q\beta}}}\right)(1\pm o(1))\right)\binom{i}{2}\right) \in \omega(n).$$

□

The same theorem holds also for $\hat{G}(Q,\hat{E})$; the proof is omitted.

6. SHRINKING/STABILIZING OF THE EFFECTIVE DIAMETER

We use the definition of the q-effective diameter given in [21].

DEFINITION 1 (EFFECTIVE DIAMETER). *For* $0 < q < 1$*, we define the* q-effective diameter *as the minimum* d_e *such that, for at least a* q *fraction of the reachable node pairs, the shortest path between the pairs is at most* d_e.

In this section we will show that the effective diameters of $G(Q,E)$ and $\hat{G}(Q,\hat{E})$ shrink or stabilize over time. The intuition behind those proofs is that even if a person q is not interested in any popular topic, and so is not linked to any popular topic in $B(Q,U)$, with high probability at least a friend of q is interested in a popular topic.

THEOREM 10. *If* $c_u < \frac{\beta}{1-\beta}c_q$*, the* q*-effective diameter of the graph* $G(Q,E)$ *shrinks or stabilizes after time* ϕn*, for any constants* $0 < \phi, q < 1$.

PROOF. Let H be the set of nodes of U in $B(Q,U)$ with degree $\geq n^\alpha$, for small $\alpha > 0$. By Lemma 2 every node in H has been inserted in the graph before time γn, for any constant $0 < \gamma < 1$. Thus the diameter of the neighborhood of H in $G(Q,E)$ shrinks or stabilizes after time γn.

Now we want to show that all but $o(n)$ nodes inserted after time ϵn with constant $\epsilon \ll \phi$ has at least a neighbor that is in the neighborhood of H in $B(Q,U)$. Hence we will be able to upper bound the q-effective diameter with the $diam(H)+2$, for any constant $q < 1$.

The number of edges that have one endpoint which is a neighbor of H is lower bounded by the number of edges generated, in the folding process, by the existence of nodes in H. At any time after ϵn the number of this edges can be lower bounded, as in Theorem 9, by
$\sum_{i=n^\alpha}^{N}\left(\left(\left(\frac{\epsilon n}{\zeta\left(-2-\frac{c_u(1-\beta)}{c_q\beta}\right)}\frac{1}{i^{2+\frac{c_u(1-\beta)}{c_q\beta}}}\right)(1\pm o(1))\right)\binom{i}{2}\right)$,
where $N = n^\gamma$, with
$\gamma < \frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}$, thus they are in $\Omega\left(n^{1+\frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}\frac{c_u(1-\beta)}{c_q\beta}}\right)$.
Instead the number of edges whose endpoints are not neighbors of H can be upper bounded by

$$\sum_{i=1}^{n^\alpha}\left(\left(\left(\frac{n}{\zeta\left(-2-\frac{c_u(1-\beta)}{c_q\beta}\right)}\frac{1}{i^{2+\frac{c_u(1-\beta)}{c_q\beta}}}\right)(1\pm o(1))\right)\binom{i}{2}\right) + sn \in \Omega\left(n^{1+\alpha\frac{c_u(1-\beta)}{c_q\beta}}\right),$$

where the first term of the sum represents all the edges that are created by nodes in U in $B(Q,U)$ and $\notin H$ and the second term represents all the edges added to the graph by a choice based on preferential attachment in $G(Q,E)$.

Now when a new node v arrives at a time between ϵn and ϕn, it chooses a set of nodes $q_{i_1},\ldots,q_{i_s}$ independently with a probability proportional to their degrees and it connects to those nodes. Thus by fixing $\alpha < \frac{1}{4+\frac{c_q\beta}{c_u(1-\beta)}}$ we have that v will point with high probability to a node that is neighbor to H in $B(Q,U)$. Hence for at least a q fraction of the reachable node pairs, the shortest path length between a pair is at most $diam(H)+2$. □

The same theorem holds also for $\hat{G}(Q,\hat{E})$; the proof is omitted.

7. SPARSIFICATION OF $G(Q,E)$

Several interesting algorithms (eg. the Dijkstra's algorithm) have complexity proportional to the number of edges in the graph. As proved in section 5, graphs produced by our model have a superlinear number of edges. In this section we will analyze two settings in which we can reduce the number of edges.

First we analyze the case in which we wish to compute shortest paths from an arbitrary node to one of several *relevant*, or distinguished, nodes. We analyze the case in which the set of relevant nodes has cardinality at most $\frac{n}{\log n}$ and is chosen uniformly at random. For this case, we present an algorithm, Algorithm $\mathscr{A}$, which, with high probability, generates from $G(Q,E)$ a new graph $G'(Q,E')$, with $|E'| \leq \delta|E|$ and $0 < \delta < 1$, such that for any node u in G and any relevant node v, a path of shortest distance in G is also present in G'.

In the second setting, in which a constant stretching of distances is allowed, we show that exists an algorithm that reduces the number of edges to a $\Theta(n)$ both in $G(Q,E)$ and in $\hat{G}(Q,\hat{E})$.

7.1 Sparsification with preservation of the distances from a set of relevant nodes

We start by describing algorithm $\mathscr{A}$, the sparsification algorithm.

Input: $G(Q,E)$ and a set R of relevant nodes.
(1) Initially, label all edges *deletable*.
(2) For each node $a \in R$:
(a) Compute the breadth first search tree starting from node a and exploring the children of a node in increasing order of insertion.
(b) Label all edges in the breadth first search tree of node a as undeletable.
(3) Delete all edges labeled as deletable.

THEOREM 11. *Suppose the set of relevant nodes R has cardinality $\frac{n}{\log n}$ and suppose that the elements of R are chosen uniformly at random from Q. If $c_u \leq \frac{\beta}{1-\beta}c_q$, the algorithm $\mathscr{A}$ with high probability generates from $G(Q,E)$ a new graph $G'(Q,E')$, with $|E'| \leq \delta|E|$ and $0 < \delta < 1$, in which the distance between every couple of nodes (a,b) is preserved if at least one of the two node is in R.*

Before proving Theorem 11 we introduce two useful lemmata.

DEFINITION 2 (USELESS NODES). *For a node $u \in U$ in $B(Q,U)$, we say that a set S_u of nodes in $G(Q,E)$ is* useless *for u if every $v \in S_u$ has an edge to u in B and, furthermore, if we compute a breadth first search in $B(Q,U)$, starting from node u and analyzing the nodes following increasing insertion order, no node in S_u will be in a path between u and a relevant node in the breadth first search tree.*

LEMMA 12. *Let $u \in U$ and let S_u be a set of useless nodes for u; then algorithm $\mathscr{A}$ will delete all edges in $G(Q,E)$ that are between nodes in S_u and that are in the clique among the neighbors of u.*

LEMMA 13. *For $\epsilon > 0$, if u has degree $\Omega(n^\epsilon)$, then a constant fraction of its neighbors will be in S_u.*

PROOF. (of Theorem 11) It is easy to see that running algorithm $\mathscr{A}$ will not change distances between pairs of nodes (a,b) if at least one of the two nodes is in R. So we have only to prove that a constant fraction of the edges are deleted by the algorithm. First we notice that we can restrict our attention only to the set F of folded edges; indeed, by construction, $|E - F| \in \Theta(n)$. Now, recalling the description of the generating process given in Theorem 9, we have that all but an $o(|E|)$ of the edges in F will be part of cliques of polynomial size generated from a node u of degree $\Omega(n^\epsilon)$, for small ϵ. Now by Lemma 12 and Lemma 13 we have that in every clique generated from such a node a constant fraction of the edges will be deleted, thus the claim follows. □

PROOF. (of Lemma 12) First we notice if an edge is deleted by $\mathscr{A}$ in $G(Q,F)$, where F is the set of folded edges, it will be deleted also in $G(Q,E)$. This is true because $\mathscr{A}$ deletes all edges that do not appear in any shortest path from any node to a node to in R and $F \subset E$. Thus in the following we will consider $G(Q,F)$.

Let $u \in U$ and $N_B(u)$ the set of neighbors of u in $B(Q,U)$. After running algorithm $\mathscr{A}$, we have that any node $v \in S_u \subset N_B(u)$ does not appear as an intermediate node in a shortest path between a relevant node and a node in $N_B(u)$. Indeed suppose by contradiction that v appears as an intermediate node in the path between a relevant node r and a node $t \in N_B(u)$, this would imply that no node $h \in N_B(u)$ would satisfy $d(h,r) \leq d(v,r)$, where $d(\cdot,\cdot)$ is the distance function, and h has been added to $B(Q,U)$ before v.
Thus the breadth first search tree in $B(Q,U)$ rooted at u should have v in the path between (u,r), thus $v \notin S_u$ a contradiction. Thus each node in S_u belongs to a different branch in every breadth first search tree in $G(Q,F)$ rooted at any relevant node, hence any edge between two nodes in S_u will be deleted by $\mathscr{A}$. □

PROOF. (of Lemma 13) As in the proof of Lemma 12 we can restict out attention to $G(Q,F)$. By Lemma 2 and Lemma 3 we have that if a node u has degree $n^\lambda \in \Omega(n^\epsilon)$ at the end of the process it should have degree μn^λ, also at time ϕn, for some constants $0 < \mu \leq 1$ and $0 < \phi < 1$, and that a constant fraction of the nodes pointing to u have been inserted after time ϕn. We call this set of nodes, L, *latecomers*. We prove that in the breadth first search from u, only $o(|L|)$ of the vertices in L are used to reach a relevant node. Thus $|S_u| \in \Theta(|L|) \in \Theta(N_B(u))$ so the Lemma will follow. In order to prove this we start by showing that the sum of the nodes over the branches of the breadth first search tree rooted at u and containing a latecomer node is $\Theta(n^\lambda)$.[5].

We say that a node i is a *child* of u if the edge (i,j) exists in $B(Q,U)$ and i has been inserted in $B(Q,U)$ after u. Let the *descendants* of u be the set S such that a node v is in S if and only if v is a child of u or v is a child of a node in S. Let E_t^{desc} be the expected number of nodes that are descendants of a *latecomer*. Notice that $E_{\phi n}^{\text{desc}} = 0$, so we have:

$$E_t^{\text{desc}} = E_{t-1}^{\text{desc}} + (\beta c_q + (1-\beta)c_u)\frac{E_{t-1}^{\text{desc}} + \mu n^\lambda}{e_{t-1} + e_{B_0}}$$

Instead of studying E_t^{desc} will we study the function W_t, with $W_{\phi n} = \mu n^\lambda$ and the recursive equation:

$$W_t = W_{t-1} + (\beta c_q + (1-\beta)c_u)\frac{W_{t-1}}{e_{t-1} + e_{B_0}}$$

[5]Note that when a node is added all its edges are copied from its prototype. So the distance between any couple of pre-existing nodes cannot shrink after the insertion of a new node. Thus in the breadth-first tree built by $\mathscr{A}$ it holds that: for a node u all the sons of u have been inserted after u

It easy to note that $W_t > E_t^{\text{desc}}$. So we have:

$$\begin{aligned}
E_t^{\text{desc}} &< W_{t-1}(1+(\beta c_q+(1-\beta)c_u)\tfrac{1}{e_{t-1}+e_{B_0}}) \\
&= W_{t-1}\left(1+\tfrac{(\beta c_q+(1-\beta)c_u)}{e_{\phi n}+(\beta c_q+(1-\beta)c_u)((t-1)-\phi n)(1\pm o(1))}\right) \\
&< W_{t-1}\left(1+\tfrac{(\beta c_q+(1-\beta)c_u)}{e_{\phi n}+(\beta c_q+(1-\beta)c_u)((1-\varphi)(t-1)-\phi n)}\right) \\
E_n^{\text{desc}} &< W_{\phi n}\tfrac{\Gamma(e_{\phi n}+(\beta c_q+(1-\beta)c_u)((1-\varphi)n-\phi n)+(\beta c_q+(1-\beta)c_u))}{\Gamma(e_{\phi n}+(\beta c_q+(1-\beta)c_u)((1-\varphi)n-\phi n))} \\
&\quad \tfrac{\Gamma(e_{\phi n}+(\beta c_q-(1-\beta)c_u)(\varphi n-\phi n))}{\Gamma(e_{\phi n}-(\beta c_q+(1-\beta)c_u)(\varphi n-\phi n)+(\beta c_q+(1-\beta)c_u))} \\
&\sim n^{\lambda}\left(\tfrac{e_{\phi n}+(\beta c_q+(1-\beta)c_u)((1-\varphi)n-\phi n)}{e_{\phi n}+(\beta c_q-(1-\beta)c_u)(\varphi n-\phi n))}\right)^{(\beta c_q+(1-\beta)c_u)} \\
&\in \Theta(n^{\lambda})
\end{aligned}$$

The final technical steps use some concentration results on *hereditary function* from the literature [14], which are omitted. Specifically, we notice that the number of descendants can be seen as a *hereditary function* on the set of edges where the boolean property is being a descendant of u. Using that the median $M[\text{number of descendants}] < c_m n^{\lambda}$ for a $0 < c_m < 1$. We have that E_t^{desc} is sharply concentrated.

Furthermore the set of relevant nodes is of cardinality $\frac{n}{\log n}$ and it is chosen uniformly at random hence with high probability only a $o(|L|)$ of the latecomers and their descendants would be a relevant. Thus only a $o(|L|)$ of the branches of the breadth first search tree rooted at u and containing a node inserted after time ϕn will lead to a relevant nodes. So all but a $o(|L|)$ of the latecomers will be in S_u.

□

7.2 Sparsification with a stretching of the distances

In the previous subsection we have shown that we can reduce the number of edges in $G(Q,E)$ by a constant factor using the algorithm $\mathscr{A}$. In this section we will study what we can achieve if we permit some bounded stretching of the shortest distance between two nodes.

We start by noticing that the graph $B(Q,U)$ has a linear number of edges and any distance between two nodes in this graph is equal to 2 times the distance of nodes in $G(Q,F)$ so adding the edges in $E-F$ it seems that we have the perfect solution to our problem. Unfortunately the original bipartite graph may not be available to us; nevertheless, we are able to explore the underlying backbone structure of G to prove the following theorem.

THEOREM 14. *There is a polynomial algorithm that, for any fixed c_u, c_q, β, finds a graph $G'(Q,E')$ with a linear number of edges, where the distance between two nodes is at most k times larger that the distance in $G(Q,E)$ and in $\hat{G}(Q,\hat{E})$, where k is a function of c_u, c_q, β.*

PROOF. First we notice that we can restrict our attention only to the folded edges, indeed by construction $|E-F| \in \Theta(n)$.

Let us say that S is a k-spanner of the graph G if it is a subgraph of G in which every two vertices are no more than k times further apart as they are in G. The problem of finding k-spanners of a graph is studied extensively in several papers — [29, 4, 5], to name a few. In our analysis, we will consider the algorithm proposed in [4] for the unit-weight case.

Their algorithm builds the set E_S of edges of the $2k$-spanner as follows: at the beginning $E_S = \emptyset$. The edges are processed one by one, and an edge is added to E_S if and only if it does not close a cycle of length $2k$ or smaller in the graph induced by the current spanner edges E_S. At the end of the process the graph $G(Q,E_S)$ will be a $2k$-spanner of $G(Q,E)$ by construction and the fact that the girth of $G(V,E_S)$ will be at least $k+1$. Since a graph with more than $n^{1+\frac{1}{k}}$ edges must have a cycle of at most $2k$ edges, the algorithm builds a spanner of size $O(n^{1+\frac{1}{k}})$.

It is important to notice that if we apply the algorithm described above to $G(Q,E)$ and $G(Q,F)$, analyzing the edges in F in the same order, every edge deleted in $G(Q,F)$ is deleted also in $G(Q,E)$. Now in the $G(Q,F)$ we have that any clique generated by any node in U has $O(n^{1+\frac{1}{k}})$ edges. Thus using the algorithm described, we have the following upper bound on the number edges for a $2k$-spanner of $G(Q,F)$.

$$|F_S| \leq \sum_{i=1}^{n}\left(\#\text{ of nodes of degree } i \text{ in } U)\left(i^{1+\frac{1}{k}}\right)\right)$$

The rest of the proof uses Theorem 1 and Lemma 4, together with some elementary manipulation, to show that $|F_S| \in \Theta(n)$. □

8. FLEXIBILITY OF THE MODEL

In this section we consider some variations of the model for which is easy to prove that the main theorems hold. We will analyze the two following cases:

- Instead of generating only one bipartite graph $B(Q,U)$, a list $B_0(Q,U),\cdots,B_k(Q,U)$ of bipartite graphs [6] are generated. At the same time the multigraph $G(Q,E)$ evolves in parallel; besides "folding" length-2 paths in $B_0,\cdots,B_k$ into edges, we also add to $G(Q,E)$ a few preferentially attached neighbors.

- Instead of "folding" length-2 paths in B into edges, for every pair of nodes in Q and every shared common neighbor $u \in U$ between them, we randomly and independently place an edge between the nodes in $G(Q,E)$ with probability proportional to the reciprocal of $d(u)^{\alpha}$, where $d(\cdot)$ denotes degree $0 < \alpha < 1$.

In the first case if for at least a bipartite graph $c_{u_i} < \frac{\beta}{1-\beta}c_{q_i}$ the densification of the edges and the shrinking/ stabilizing follow using the same arguments used in the proof of the theorems 9 and 10. Furthermore if k is constant all the theorems on the degree distribution of $G(Q,E)$ and $\hat{G}(Q,\hat{E})$ continue to hold.

In the second case it is sufficient to notice that every node u in U in $B(Q,U)$ is no longer substituted by a clique but by a $G(n,p)$, where $n = d(u)$ and $p = \frac{1}{d(u)^{\alpha}}$. Now if $c_u < \frac{\beta}{1-\beta}c_q(1-\alpha)$ using the same argument of Theorem 5 and the Chernoff bound we obtain the densification of the edges. The shrinking/stabilizing diameter in this case follows from the fact that most of the nodes will point to a high degree node in $G(Q,E)$[7] and that the $G(n,p)$, where $p = n^{\alpha}$ for $0 < \alpha < 1$, has constant diameter by [8]. Finally also in this case the degree distribution is heavy-tailed because with high probability the complementary cumulative distribution

[6] In this model the choice of adding a node to U or Q is the same for all the graphs, but the number of edges added$(c_{u_0}, c_{q_0}, \cdots, c_{u_k}, c_{q_k})$ and their destination differ.

[7] This can be proved using the same proof strategy as before.

function of nodes dominates the complementary cumulative distribution function of the degrees of Q in $B(Q, U)$.

Acknowledgments

We are grateful to Patrick Nguyen (U. British Columbia and Google) for joint experimental work that greatly guided our thinking about these ideas, and to Mark Sandler (Google) for numerous insightful comments and discussions.

9. REFERENCES

[1] W. Aiello, F. Chung, L. Lu, "Random Evolution in Massive Graphs." in 42nd IEEE symposium on Foundations of Computer Science, FOCS 2001.

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

[3] N.Alon, S. Hoory and N. Linial. "The moore bound for irregular graphs." in Graphs and Combinatorics 18 (2002), pages 53-57.

[4] I. Althófer, G. Das, D. Dobkin, D. Joseph, and J. Soares, "On sparse spanners of weighted graph." in Discrete and Computational Geometry, 9:81-100, 1993.

[5] S.Baswana, S. Sen, "A simple linear time algorithm for computing sparse spanners in weighted graphs." in 30th International Colloquium on Automata, Languages and Programming(ICALP), pages 384-396, 2003.

[6] B. Bollobás, O. Riordan, J. Spencer, and G. E. Tusnády. "The degree sequence of a scale-free random graph process." Random Structures and Algorithms, 18(3):279-290, 2001.

[7] B. Bollobás, O. Riordan, "The Diameter of a Scale-Free Random Graph." Combinatorica 24(1): 5-34 2004.

[8] B. Bollobás, "The Diameter of Random Graphs." IEEE Trans. Inform.Theory 36 (1990), no. 2, 285-288.

[9] R. L. Breiger "The Duality of Persons and Groups." Social Forces, University of North Carolina Press, 1974.

[10] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. "Graph structure in the web." In Proc. 9th International World Wide Web Conference, pages 309-320, 2000.

[11] F. Chierichetti, S. Lattanzi and A. Panconesi, "Gossiping in Social Networks." to be submitted.

[12] C. Cooper and A. M. Frieze. "A general model of web graphs." Random Structures and Algorithms, 22(3):311-335, 2003.

[13] M. Dodds and Watts. "An experimental study of search in global social networks." Science, 301(5634):827-829, 2003.

[14] D. Dubhashi, "Talagrand's inequality in hereditary settings." in Technical report, Dept. CS, Indian Istitute of Technology, 1998.

[15] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. "Heuristically optimized trade-offs: A new paradigm for power laws in the internet." In Proc. 29th International Colloquium on Automata, Languages and Programming, pages 110-122, 2002.

[16] M. Faloutsos, P. Faloutsos, C. Faloutsos. "On Power-law Relationships of the Internet Topology." SIGCOMM 1999: 251-262.

[17] J. Kleinberg. "Navigation in a small world." Nature, 406:845, 2000.

[18] J. Kleinberg. "The small-world phenomenon: An algorithmic perspective." In Proc. 37th Annual ACM Symposium on Theory of Computing, pages 163-170, 2000.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. "Stochastic models for the web graph." In Proc. 41st IEEE Symposium on Foundations of Computer Science, pages 57-65, 2000.

[20] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins "Microscopic evolution of social networks." KDD 2008:462-470.

[21] J. Leskovec, D. Chakrabarti, J.M. Kleinberg, C. Faloutsos, "Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication." PKDD 2005: 133-145.

[22] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2005.

[23] M. Mahdian, Y. Xu, "Stochastic Kronecker Graphs." WAW 2007: 179-186.

[24] B. Mandelbrot. "An informational theory of the statistical structure of languages." In W. Jackson, editor, Communication Theory, pages 486-502. Butterworth, 1953.

[25] C.J.H. McDiarmid "On the method of bounded differences." In J. Siemons editor, Surveys in Combinatorics: Invited Papers at the 12th British Combinatorial Conference, number 141 in London Mathematical Society Lecture Notes Series, pages 148-188. Cambridge University Press, 1989.

[26] M. Mitzenmacher. "A brief history of generative models for power law and lognormal distributions." Internet Mathematics, 1(2), 2003.

[27] M. Mitzenmacher. "Editorial: The Future of Power Law Research." Internet Mathematics, vol. 2. no. 4, pp. 525-534, 2006.

[28] D. L. Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins "Geographic Routing in Social Networks." Proceedings of the National Academy of Sciences, Vol. 102, No. 33. (August 2005), pp. 11623-11628.

[29] D. Peleg and E. Upfal, "A trade-off between space and efficiency for routing tables." in Journal of Assoc. Comp. Mach., 36(3):510-530, 1989.

[30] H. Simon. "On a class of skew distribution functions." Biometrika, 42:425-440, 1955.

[31] J. Travers and S. Milgram. "An experimental study of the small world problem." Sociometry, 32(4):425-443, 1969.

[32] D. Watts and S. Strogatz. "Collective dynamics of small-world networks." Nature, 393(6684):409-410, 1998.

[33] G. K. Zipf. "Human Behavior and the Principle of Least Effort." Addison-Wesley, 1949.