An Incremental Super-Linear Preferential Internet Topology Model

Extended Abstract

Sagy Bar[1], Mira Gonen[2], and Avishai Wool[3]

[1] School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978, ISRAEL. sagyb@eng.tau.ac.il.

[2] Dept. Computer Science, Tel Aviv University, Ramat Aviv 69978, ISRAEL. gonenmir@tau.ac.il.

[3] School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978, ISRAEL. yash@acm.org.

Abstract. By now it is well known that the distribution of node degrees in the graph induced by the peering arrangements between Autonomous Systems (ASs) exhibits power laws. The most appealing mathematical model that attempts to explain the power-law degree distribution was suggested by Barabási and Albert (the BA model). We introduce two new models that are extensions to the BA model: the "Incremental Edge Addition" (InEd) model, and the "Super-Linear Preferential Attachment" (SLiP) model. We prove that both our models are more successful in matching the power-law exponent, in producing leaves , and in producing a large dense core. Beyond mathematical analysis, we have also implemented our models as a synthetic network generator we call Tang (Tel Aviv Network Generator). Experimentation with Tang shows that the networks it produces are more realistic than those generated by other network generators.

1 Introduction

1.1 Background and Motivation

The connectivity of the Internet crucially depends on the relationships between thousands of Autonomous Systems (ASes) that exchange routing information using the Border Gateway Protocol (BGP). These relationships can be modeled as a graph, called the AS-graph, in which the vertices model the ASes, and the edges model the peering arrangements between the ASes.

Significant progress has been made in the study of the AS-graph's topology over the last few years. In particular, it is now known that the distribution of vertex degrees (i.e., the number of peers that an AS has) is heavy-tailed and obeys so-called power-laws [SFFF03]: The fraction of vertices with degree k is proportional to $k^{-\gamma}$ for some fixed constant γ. This phenomenon cannot be explained by traditional random network models such as the Erdős-Renyi model [ER60].

5th Annual Passive & Active Measurement Workshop (PAM), Antibes Juan-les-Pins, France, April 2004.

1.2 Related Work

Barabási and Albert [BA99] introduced a very appealing mathematical model to explain the power-law degree distribution (the BA model). The BA model is based on two mechanisms: (i) networks grow incrementally, by the adding new vertices, and (ii) new vertices attach preferentially to vertices that are already well connected. They showed, *analytically*, that these two mechanisms suffice to produce networks that are governed by a power-law.

While the pure BA model [BA99] is extremely elegant, it does not accurately model the Internet's topology in several important aspects:

- The BA model does not produce any leaves (vertices with degree 1), whereas in the real AS-graph some 30% of the vertices are leaves.
- The BA model predicts a power law distribution with a parameter $\gamma = 3$, whereas the real AS-graph has a power law with $\gamma \approx 2.11$. This is actually a significant discrepancy: For instance, the most connected ASes in the AS graph have 500–2500 neighbors, while the BA model predicts maximal degrees which are roughly 10 times smaller on networks with comparable sizes.
- It is known that the Internet has a surprisingly large *dense core* [SARK02], [SW03a]: The AS graph has a core of 43 ASes, with an edge density ϱ[1] of over 70%. However, as recently shown by Sagie and Wool [SW03b], the BA model is fundamentally *unable* to produce synthetic topologies with a dense core larger than $\ell = 6$ with $\varrho(\ell) \geq 70\%$.

These discrepancies, and especially the fact that the pure BA model produces an incorrect power law parameter $\gamma = 3$, were observed before. Barabási and Albert themselves refined their model in [AB00] to allow adding links to existing edges, and to allow rewiring existing links. However, as argued by [CCG$^+$02], the idea of link-rewiring seems inappropriate for the AS graph, and [BT02] showed that the rewiring probability needs to be as high as 50% for the [AB00] model to produce values of γ which are closer to reality.

The work closest to ours is that of Bu and Towsley [BT02]. The authors attempted to produce a BA-like model that will (i) produce a more realistic power law parameter γ, while (ii) still remaining amenable to mathematical analysis. The model they suggest meets these goals, however, we claim that it is still not satisfactory. Their model is rather unnatural, and involves a crucial technical parameter that does not correspond to any intuitive feature of the development of the AS graph. The authors themselves admit that the main reason for considering such a counter-intuitive model is that it can be analyzed mathematically—and that other, more natural, extensions, greatly increase the difficulty of analysis.

[1] The density $\varrho(\ell)$ of a subgraph with ℓ vertices is the fraction of the $\ell(\ell-1)/2$ possible edges that exist in the subgraph.

1.3 Contributions

Our main contribution is a new extension to the BA model, that has the following features:

- It addresses the discrepancies of the BA model with respect to (i) the lack of leaves, (ii) value of the power law parameter γ, and (iii) the lack of a dense core.
- It is natural and intuitive, and follows documented and well understood phenomena of the Internet's growth.
- We are able to analyze our model, and rigorously prove many of its properties.

Our model hinges on two ideas, which we call "Incremental Edge Addition" (InEd) and "Super-Linear Preferential Attachment" (SLiP)

Beyond mathematical analysis, we also implemented our model as a synthetic network generator we call TANG (Tel Aviv Network Generator). Experimentation with TANG shows that the networks it produces are more realistic than those generated by other network generators such as BRITE [MLMB01], and Inet [WJ02]. TANG is freely available from the authors [Woo04].

Organization: In the next section we give an overview of the BA model. In Section 3 we introduce the Incremental Edge Addition (InEd) model. Section 4 presents the Super-Linear Preferential Attachment (SLiP) model. In Section 5 we analyze the expected number of leaves in a model combined from the InEd model and the SLiP model. Section 6 describes TANG and the results of our simulations. We conclude with Section 7.

We omit most of the proofs in this space-limited extended abstract.

2 Overview of the BA Model

The BA model works as follows. (i) Start with a small number (m_0) of arbitrarily connected vertices. (ii) At every time step, add a new vertex with $m(\leq m_0)$ edges that connect the new vertex to m different vertices already present in the system. (iii) The new vertex picks its m neighbors randomly, where an existing vertex i, with degree k_i, is chosen with probability $p(k_i) = k_i / \sum_j k_j$.

Since every time step introduces 1 vertex and m edges, it is clear that the average degree of the resulting network is $\approx 2m$.

Observe that new edges are added in *batches* of m. This is the reason why the pure BA model never produces leaves, [SW03a], and the basis for the model's inability to produce a dense core. Furthermore, empirical evidence [CCG+02] shows that the vast majority of new ASes are born with a degree of 1, and not 2 or 3 (which is necessary to reach the AS graph's average degree of ≈ 4.33).

3 The Incremental Edge Addition (InEd) Model

Our first model modifies the way in which edges are introduced into the BA model. In this section we give the model's definition, analyze its degree distribution and prove that it is close to a power-law distribution. We also analyze the expected number of leaves.

3.1 Model Definition

The basic setup in the InEd model is the same as in the BA model: We start with m_0 nodes. At each time step we add a new node, and m edges. However, the edges are added in the following way: one edge connects the new node to nodes that are already present. An existing vertex i, with degree k_i, is chosen with probability $p(k_i) = k_i / \sum_j k_j$. (That is, $p(k_i)$ is linear in k_i, as in the BA model). The remaining $m-1$ edges connect *existing* nodes. One endpoint of each edge is uniformly chosen, and the other endpoint is connected preferentially, choosing a node i with probability $p(k_i)$ as defined above.

Note that this is reminiscent of the [AB00] model. In that model nodes are all added with degree m, and additionally, nodes that are chosen uniformly at random grow more edges with some fixed probability p. In our model, all nodes start with degree 1, as found empirically by [CCG$^+$02]. Moreover, we avoid the extra parameter p.

Our analysis shows that the InEd model produces a remarkably accurate number of leaves, and a power-law degree distribution, albeit with a parameter γ which is still too high. The predicted maximal degree improves as well: it is about twice that predicted by the BA model.

3.2 Power Law Analysis

We show that the InEd model produces a near-power-law degree distribution. We analyze our model using the "mean field" methods in Barabási-Albert [BA99]. As in [BA99], we assume that k_i changes in a continuous manner, so k_i can be interpreted as the average degree of node i, and the probability $p(k_i)$ can be interpreted as the rate at which k_i changes.

Theorem 3.1. *In the InEd model,* $\Pr[k_i(t) = k] \propto (k + 2m - 2)^{-3}$.

We prove the theorem using the following lemma.

Lemma 3.2. *Let t_i be the time at which node i was added to the system. Then* $k_i(t) = (2m-1)\sqrt{\frac{t}{t_i}} - 2(m-1)$.

Proof: At time t the sum of degrees is $2mt$. The change in an existing node's degree is influenced by the probability of it being chosen preferentially, and by the probability that it is selected uniformly. Thus we get the following differential equation:

$$\frac{\partial k_i}{\partial t} = m \cdot \frac{k_i}{2mt} + \frac{m-1}{t} = \frac{k_i}{2t} + \frac{m-1}{t}.$$

The initial condition for node i is $k(t_i) = 1$. Solving for $k_i(t)$ proves the Lemma. ∎

Corollary 3.3. *The expected maximal degree in the InEd model is*

$$(2m-1)(\sqrt{t}-1)+1.$$

Proof: By setting $t_i = 1$ in Lemma 3.2 we get the result. ∎

Proof of Theorem 3.1: Using Lemma 3.2 the probability that a node has a degree $k_i(t)$ smaller than k, $Pr[k_i(t) < k]$, can be written as

$$\begin{aligned}\Pr\left[k_i(t) < k\right] &= \Pr\left[(2m-1)\sqrt{\frac{t}{t_i}} - 2(m-1) < k\right] = \Pr\left[t_i > \left(\frac{2m-1}{k+2m-2}\right)^2 t\right]\\ &= 1 - \Pr\left[t_i \le \left(\frac{2m-1}{k+2m-2}\right)^2 t\right] = 1 - \left(\frac{2m-1}{k+2m-2}\right)^2 \frac{t}{t+m_0}\end{aligned}$$

Thus

$$\Pr\left[k_i(t) = k\right] = \frac{\partial}{\partial k}\left[1 - \left(\frac{2m-1}{k+2m-2}\right)^2 \frac{t}{t+m_0}\right] \propto (k+2m-2)^{-3} \qquad ∎$$

Theorem 3.1 shows that the InEd model produces a near-power-law distribution, but the coefficient γ is still ≈ 3.

3.3 Analysis of the Expected Number of Leaves

The pure BA model is unable to produce any leaves: each new node has degree m. In contrast, the InEd model produces a realistic number of leaves. Note that nodes in the InEd model start as leaves. We now compute the probability that a node that entered at time t_i will remain a leaf at time n, and compute the expected number of leaves in the system at time n.

Let v_i be the node that entered at time t_i, and let $deg_n(v_i)$ be the degree of v_i after time n.

Theorem 3.4. *In the InEd model,* $E[\#leaves] \le \frac{n}{m+1/2}$.

Computer simulations show that this upper bound is very accurate: for $n = 10,000$, $m = 2$, the bound of Theorem 3.4 is 40% leaves, and our simulation show that about 3,995 leaves are generated.

4 The Super-Linear Preferential Attachment (SLiP) Model

In this model, we generalize the BA model in a different way: We assume that the utility of joining a highly-connected node is super-linear in its degree. This assumption agrees with the observations of [CCG+02]. As in Section 3, we give the model's definition, analyze its degree distribution and prove that it is close to power-law distribution.

4.1 Model Definition

In the SLiP model, at each time step we add a new node, and m edges, in the following way: All m edges connect the new node to nodes already present in the network (as in the pure BA model). However, an existing node i is chosen as an endpoint with probability

$$p(k_i) = \frac{k_i^{1+\varepsilon}}{\sum\limits_j k_j^{1+\varepsilon}},$$

for some $\varepsilon > 0$. Thus the preferential attachment is super linear. Note that setting $\varepsilon = 0$ gives the pure BA model.

4.2 Power Law Analysis

As in the analysis of the InEd model, we show that the SLiP model produces a near-power-law distribution. As before we assume that k_i changes in a continuous manner, so the probability $p(k_i)$ can be interpreted as the rate at which k_i changes.

A main technical difficulty in the SLiP model is that the denominator $\sum k_j^{1+\varepsilon}$ is not fixed. Therefore, we start by bounding $\sum\limits_j k_j^{1+\varepsilon}$.

Lemma 4.1. *For any network over t nodes and mt edges, and any $\varepsilon > 0$,*

$$t(2m)^{1+\varepsilon} \leq \sum_j k_j^{1+\varepsilon} \leq (2mt)^{1+\varepsilon}$$

Corollary 4.2. $\sum_j k_j^{1+\varepsilon} \approx (2m)^{1+\varepsilon} t^{1+\varepsilon/2}$

Lemma 4.3. *In the SLiP model,* $k_i(t) = m \Big/ \left(1 - \frac{1}{2^\varepsilon t_i^{\varepsilon/2}} + \frac{1}{2^\varepsilon t^{\varepsilon/2}}\right)^{1/\varepsilon}$

Corollary 4.4. *The expected maximal degree in the SLiP model is* $\leq 2m\sqrt{t}$

Corollary 4.4 shows that the SLiP model, on its own, achieves essentialy the same (expected) maximal degree that is achieved by the InEd model (recall Corollary 3.3). This maximal degree is about twice higher than that of the pure BA model.

Theorem 4.5. *In the SLiP model* $\Pr[k_i(t) < k] = 1 - \left[\frac{(1/t+m_0)^{\varepsilon/2}}{2^\varepsilon + \frac{1}{t^{\varepsilon/2}} - \left(\frac{2m}{k}\right)^\varepsilon}\right]^{2/\varepsilon}$

Note that the SLiP model does not produce any leaves since nodes are added with degree m.

5 The Combined InEd/SLiP Model

Since the InEd and SLiP models modify the BA model in different ways, we can easily combine them into a single model, which would enjoy the benefits offered by each model. Unfortunately, we are unable to show, analytically, that the combined model has a power-law behavior—the differential equations we obtain are too difficult.

5.1 Analysis of the Expected Number of Leaves

In contrast, we are able to analyze the expected number of leaves in the combined model. Theorem 5.1 shows that the bound of Theorem 3.4 almost holds for the combined model as well, up to a small constant factor.

As in the InEd Model, let v_i be the node that entered at time t_i, and let $deg_n(v_i)$ be the degree of v_i after time n.

Theorem 5.1. *In the SLiP model,* $E[\#leaves] \leq \frac{n}{m}$.

6 Implementation

We implemented the combined SLiP/InEd model as a synthetic network generator we call Tang (Tel Aviv Network Generator). Tang accepts the desired number of vertices (n), the average degree (d), and the utility function's exponent ($a = 1 + \varepsilon$), as arguments. The average degree is allowed to be fractional. Setting the exponent to 1 (i.e., $\varepsilon = 0$) causes Tang to use the linear InEd model. Tang is also able to produce pure BA-model networks.

We used Tang to generate synthetic topologies with Internet-like parameters. We used $n = 15,000$ and $d = 4.33$, which match the values reported in [SW03a]. We generated 10 random topologies for each setting of $\varepsilon = 0, 0.1, 0.2, 0.3$, and 10 random topologies for the pure BA model. We compared these networks to the AS-graph snapshot collected by [SW03a].

6.1 Power Law Analysis

Fig. 1 shows the Complementary Cumulative Density Function (CCDF)[2] of the degree distribution in the Internet's AS-graph and in the Tang-generated synthetic networks. For the synthetic networks, each CCDF curve is the average taken over the 10 randomly generated networks.

The figure clearly shows that the AS graph obeys a power-law degree distribution, with a CCDF exponent of $\eta = 1.17$. The figure also shows the shortcomings of the pure BA model: (a) we can see that $CCDF(2) = CCDF(1) = 1$, which indicates that BA networks do not contain any leaves; and (b) it is clear that slope of the BA model's CCDF is too steep: the power-law exponent is $\eta = 1.96$.

[2] For any distribution of degrees, $CCDF(k) = \Pr[deg_n(v) \geq k]$. Note that if $\Pr[deg_n(v) = k] \propto k^{-\gamma}$ then $CCDF(k) \propto k^{-\eta} = k^{1-\gamma}$.

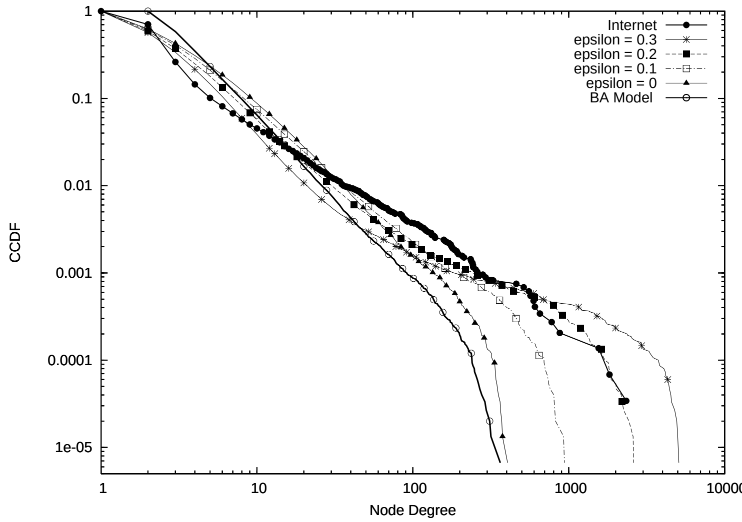

Fig. 1. The CCDF of the degree distribution for the Internet's AS-graph, the combined SLiP/InEd networks with $\varepsilon = 0, \ldots, 0.3$, and the pure BA model (log-log scale).

The figure shows that the InEd model ($\varepsilon = 0$) brings the number of leaves in the network to a fairly realistic level: 37.5% leaves in the InEd model versus 30% in the AS-graph. Note that Theorem 3.4 predicts that when the average degree is 4.33 (i.e., $m = 2.165$) the number of leaves will be $1/(2.165 + 0.5) = 37.52\%$: a very accurate estimate. We can see that the power law produced by the InEd model is slightly better than that of the BA model ($\eta = 1.83$), but still too steep.

The figure shows that the SLiP model shifts the CCDF curve closer to the Internet curve as ε grows to 0.1 and 0.2. However, when ε reaches 0.3 the CCDF overshoots the Internet curve in the high-degree area (above 800 neighbors), and undershoots the Internet curve in the mid range (10-800 neighbors). This "S" shape becomes even more pronounced with $\varepsilon = 0.4$ or higher (curves omitted). Intuitively, the SLiP model makes the high-degree nodes more attractive at the expense of low- and mid-degree nodes, and setting ε too high amplifies this behavior beyond what is observed in reality. We can see that the networks with the most realistic degree distribution are generated with $\varepsilon = 0.2$, in which case the power-law exponent is $\eta = 1.13$.

6.2 Dense Core Analysis

In order to find the Dense Core in the networks, we used the Dense k-Subgraph (DkS) algorithms of [FKP01,SW03a]. These algorithms search for the densest cluster (sub-graph) of a prescribed size ℓ. Fig. 2 shows the edge density of the densest cluster found by the algorithms, as a function of ℓ. For the synthetic

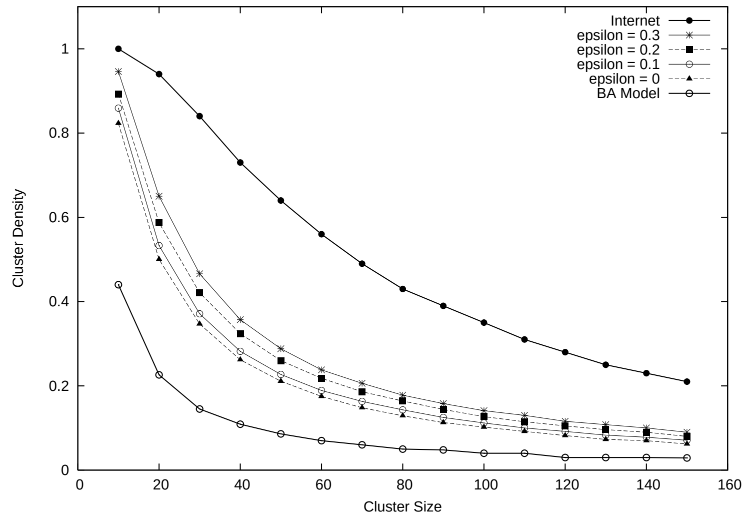

Fig. 2. The edge density $\varrho(\ell)$ of the densest ℓ-cluster, as a function of the cluster size ℓ.

networks, each point on the curves is the average over 10 random networks generated with the same parameters.

The figure clearly shows that for Internet-like parameters, the pure BA model does not produce significant a dense core: This is not surprising in view of the results of Sagie and Wool [SW03b], who proved that the BA model is fundamentally *unable* to produce synthetic topologies with a dense core larger than $\ell = 6$ with $\varrho(\ell) \geq 70\%$. In contrast, the real AS graph has a dense core of $\ell = 43$ ASes with $\varrho(\ell) \geq 70\%$.

The figure does show that the TANG-generated networks have dense cores that are closer to reality than those produced by the pure BA model: we see that a density of $\varrho(\ell) \geq 70\%$ is achieved around $\ell \in [17, 20]$, and that higher values of ε produce larger dense cores. In fact, for any value of ℓ, the density $\varrho(\ell)$ of the TANG-generated networks is at least twice the density of the BA networks. Thus, as far as dense clusters go, TANG is significantly closer to reality than the BA model.

However, the figure also shows that dense cores of TANG networks still fall short: they are roughly half as dense as their counterparts in the AS graph. Furthermore, increasing ε only produces a slow increase in the density of the core, and we already saw in Section 6.1 that increasing ε beyond 0.2 distorts the degree distribution away from a power law. Thus, we conclude that the SLiP/InEd model is a significant improvement in terms of the dense core—but it is not sufficient to produce realistic cores.

7 Conclusions and Future Work

We have shown that our extensions to the BA model, the InEd and SLiP models, significantly improve upon the pure BA model in terms of matching the power-law parameter, producing leaves, and producing a large dense core. Our models are amenable to mathematical analysis, and are implemented as a freely available network generator.

However, more work is possible: The current model still does not produce a satisfactory dense core. It seems that new ideas are necessary to create a model that can (i) produce larger dense cores, (ii) maintain a power law degree distribution, and (iii) remain simple enough to analyze.

References

[AB00] Réka Albert and Albert-László Barabási. Topology of evolving networks: Local events and universality. *Physical Review Letters.*, 85(24):5234–5237, December 2000.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 15 October 1999.

[BT02] T. Bu and D. Towsley. On distinguishing between Internet power-law generators. In *Proc. IEEE INFOCOM'02*, New-York, NY, USA, April 2002.

[CCG+02] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. In *Proc. IEEE INFOCOM'02*, New-York, NY, USA, April 2002.

[ER60] P. Erdős and A. Renyi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 5:17–61, 1960.

[FKP01] U. Feige, G. Kortsarz, and D. Peleg. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[MLMB01] A. Medina, A. Lakhina, I. Matta, and J. Byers. BRITE: An approach to universal topology generation. In *Proceedings of MASCOTS'01*, August 2001.

[SARK02] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. In *Proc. IEEE INFOCOM'02*, New-York, NY, USA, April 2002.

[SFFF03] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power-laws and the AS-level Internet topology, 2003. To appear in IEEE/ACM Trans. on Networking.

[SW03a] G. Sagie and A. Wool. A clustering approach for exploring the Internet structure. Technical Report EES2003-7, Dept. Electrical Engineering Systems, Tel Aviv University, 2003. Available from `http://www.eng.tau.ac.il/~yash/ees2003-7.ps`.

[SW03b] G. Sagie and A. Wool. A clustering-based comparison of Internet topology models, 2003. Preprint.

[WJ02] Jared Winick and Sugih Jamin. Inet-3.0: Internet topology generator. Technical Report UM-CSE-TR-456-02, Department of EECS, University of Michigan, 2002.

[Woo04] A. Wool. TANG: Tel aviv university network generator, v1.4, 2004. Available from `http://www.eng.tau.ac.il/~yash/topology/index.html`.