Comparing Trade-off Based Models of the Internet

Anthony Spatharis*, Ilias Foudalis, Martha Sideri

Computer Science Department
Athens University Of Economics And Business
Patision 76, 10434 Athens, Greece
spathari@aueb.gr, fountalis@aueb.gr, sideri@aueb.gr

Christos Papadimitriou
Computer Science Department
University of California
Berkeley, USA
christos@cs.berkeley.edu

Abstract. We introduce and evaluate several new models of network growth. Our models are extensions of the FKP model, modifying and improving it in various dimensions. In all these models nodes arrive one by one, and each node is connected to previous nodes by optimizing a trade-off between a geometric objective (“last mile cost”) and a topological objective (“position in the network”). Our new models differ from the original FKP model in directions inspired by the real Internet: two or more edges are attached to each arriving node (while the FKP model produces a tree); these edges are chosen according to various criteria such as robustness; edges may be added to the network between old nodes; or only certain “fertile” nodes (an attribute that changes dynamically) are capable of attracting new edges. We evaluate these models, and compare them with the graph of the Internet’s autonomous systems, with respect to a suite of many test parameters (such as average degree, power law exponent, and local clustering rank) proposed in the literature; to this end we have developed the network generation and measurement system Pandora.

Keywords: Internet topology, power law, scale-free graphs, network generators, network modeling and simulation, complex networks

*Address for correspondence: Department of Computer Science, Athens University of Economics and Business, Patission 76, 10434 Athens, Atiki, Greece
1. Introduction

The Internet is a worldwide, publicly accessible collection of many thousands of interconnected computer networks; it is also the platform of a significant and increasing part of humanity’s economic and social activity. Understanding the structure of the Internet graph (the graph of the routers, or that of the 20,000 or so autonomous systems comprising the Internet), is a major research frontier in computer science. In particular, it is of great importance to develop realistic generative models of Internet-like graphs, since proposed new protocols and algorithms need to be tested on models that are as realistic as possible. In this regard, the observation, first made in [14], that the degrees of the Internet graph are power law distributed (instead of the normal distribution predicted by traditional random graph models) brought about a revolution in the area, ushering in sophisticated graph models intended to model Internet growth. Indeed, starting with the preferential attachment model [3], many such models have been proposed; see [1, 6, 11, 16, 17] for some surveys, and see [5, 7, 10, 12, 18, 21] for several reviews and bibliographies on scale-free networks from different perspectives.

One of the most influential models of network growth is the so called FKP model proposed in [13]. In this model, nodes arrive in discrete times, one after the other, and each node is connected to the node that minimizes a linear combination of two objectives; euclidean distance, and topological (hop) distance from the center of the network (assumed to be the first node). While this model has been shown in [13] to produce power law distributed degrees for large areas of the trade-off, it has also several shortcomings. For example, the produced network is a tree, and in fact one with a huge number of leaves; as a consequence the average degree is two (much less than the average degree of the real Internet). See [4] for a comprehensive critique of the FKP model.

In this paper we extend the FKP model in several novel directions, addressing these and other shortcomings. The models proposed differ from the original FKP model in various respects. In all but one of the models each new node is connected to the network via not one but several edges, chosen so to optimize the original trade-off; as a result, the network is no longer a tree. How these edges are chosen among those that have good trade-off values defines several variants of our model. One may insist that the few best edges are added; or that two edges that are, in some precise sense, “far” from one another. In some models an edge may be added, with some small probability, between nodes of the network other than the one that has just arrived; this reflects real-life situation in which new connections force an already existing node to increase its connectivity to the rest of the Internet. Finally trying to model another real-life phenomenon, we designate some nodes as fertile, an attribute which changes dynamically. New nodes can only connect to nodes that are presently fertile. In a recent authoritative survey of the area [17] it is pointed out that, when it comes to models of the Internet, there is a trade-off between simplicity and realism; that is, how accurately the model reflects properties of the real Internet, and how complicated and unintuitive does it have to be to achieve such closeness. We believe that the models presented here occupy a very attractive region in this trade-off.

Importantly, we test these new models experimentally by calculating several key metrics proposed in the literature, such as average degree, power-law degree, clustering coefficient, etc. We use these metrics to compare the various models, one against the other as well as each with previously proposed models, and with the real Internet. This comprehensive testing methodology, as well as the experimental system Pandora for performing these tests that we have created and made available on the Internet, is another contribution of this paper. Our experiments show that these models produce graphs that are in excellent agreement with the Internet with respect to the test parameters.
The rest of the paper is organized as follows. Section two provides a brief review of the background and related work. In section three the new models are presented and discussed. In section four we present our experimental results, measurements, and comparisons, and we briefly describe the Pandora system. Finally, in section five we discuss our findings and chart future work.

2. Background and Related Work

The oldest and most influential model of random graphs is the 1959 $G[n, p]$ model of Erdős and Rényi, in which edges among $n$ nodes are selected independently with probability $p$; the degrees are thus concentrated by a Gaussian distribution, according to the central limit theorem. However, it was observed in [14] that the degree distribution of the Internet topology exhibits a power law degree distribution:

$$P(k > x) \propto x^{-\gamma},$$

(1)

where $\gamma$ is the shape parameter or tail exponent. It was later pointed out that, in addition, Internet graphs are highly clustered, among other properties. Many evolutionary models (in which nodes arrive one by one and are somehow connected to the already existing nodes) have been proposed to explain these observations. Of these, the preferential attachment models assume that the arriving node is connected to old ones with probability proportional to their degrees [1, 3]. In the trade-off based models [2, 9, 13] the connections are such that a multi-objective criterion is optimized. Both of these families of models have been shown to generate power law degree distributions, but are lacking with respect to other criteria. We next focus on the FKP model.

The original trade-off model FKP [13] is a simple evolutionary model. Each node $i$ (after the first one) is a uniformly random point in the unit square, and attaches itself to the node $j$ that minimizes this weighted sum of two objectives:

$$\text{min}_{j<i}(\alpha * d_{ij} + ecc(j)).$$

(2)

The first term is the Euclidean distance between the two nodes (the “last mile cost”), while the second is the eccentricity of $j$, its distance from the center (assumed for simplicity to be the first node). $\alpha$ is a weight, capturing the relative importance of the two objectives. It is proved in [13] that, for all but extremely small and extremely large values of this parameter, the degree sequence of the resulting graph is power law distributed [4, 13]. The conceptual advantage of the FKP model over, say, the preferential attachment model is that attachment preference is based on a plausible network-motivated objective, as opposed to being postulated.

3. The Network Models

There are many reasons why, despite its initial promise, the original FKP model fails to be a good model of the Internet Topology. Since each node attaches itself to the network by a single edge, the resulting graph is a tree; to put it otherwise, the average degree of each node is two, whereas the average degree of the Internet AS graph is 4.6.
3.1. Best Two (BT) model

In the simplest and most basic modification of FKP called Best Two, when a new node \( i \) arrives, instead of selecting one node \( j \) that minimizes (2), the two best such nodes are chosen, and are connected to node \( i \). This immediately gives the graph a nontree structure and increases the average degree to four, much closer to the Internet’s 4.6. One disadvantage of this method is that the two nodes that optimize (2) may be very close, resulting in connections and paths of low diversity.

![Figure 1](image1)

Figure 1. Graph generated by the best two algorithm. In all figures, node labels are order of arrival; \( \alpha \) is assumed to be 2.

For example, in Figure 1 when the new node 5 arrives, the FKP criterion first chooses node 2, while the second node chosen will be node 3.

3.2. Independent path (IP) model

In this model when the new node \( i \) arrives, first is linked to a node \( j \) according to (2). A second edge is attached from the new node \( i \) to another node \( k \neq j \) that minimizes the following quantity:

\[
\min_{k < i, k \neq j} (\alpha \ast d_{ik} + \text{Com}(j, k)),
\]

where \( \text{Com}(j, k) \) is the number of common nodes of the two shortest paths from \( j \) and \( k \) to the center. In other words, this model favors second connections that are different from the first.

![Figure 2](image2)

Figure 2. Graph generated by the independent path algorithm.
Let us see what happens when node 5 enters the graph in Figure 2. The first node selected is node 2. But in contrast with the BT model, the second node selected is node 4.

3.3. Enhanced models

In the real Internet, connections are not created only when a new node arrives. As old nodes acquire more connections and attract more traffic they may seek new upstream connections. This is captured by adding to the previous models the following additional step:

After the new node has been connected to the two previous nodes, one of these two previous nodes, say \( j \), is connected, with probability \( p \), to another existing node, called the peering node. The peering node is the one which minimizes (2) for. By adding this feature to the Best of Two and the Independent Path models, one obtains the Enhanced Best Two (EBT) and Enhanced Independent Path (EIP). When \( p = 0 \) the non-enhanced models are obtained, while for \( p = 0.3 \) the average degree of the resulting graph is 4.6, the same as of that of the real Internet.

3.4. The controlled distance (CD) model

In this new model, when a new node \( i \) arrives, it connects to an existing node \( j \) in the graph, that minimizes (2). A second edge connects node \( i \) to an existing node \( k \) that minimizes (2), subject to the constraint that the hop distance between nodes \( j \) and \( k \) is at least a fixed lower bound, such as 4. To implement this algorithm one needs a data structure that maintains the hop distance between any two nodes in the graph.

For example, in Figure 3 the second node selected by node 4 could be node 1 and not node 2 (as would happen in the BT model, for example), because node 1 has larger hop distance from the first connection, node 3.

3.5. The Fertile Node (FN) model

The graph of the real Internet has a low power law exponent, and only 26% of its nodes as leaves. The FKP model also has a low power law exponent, but the great majority of its nodes are leaves. The Fertile Node (FN) model achieves a low exponent while at the same time maintaining a low number of leaves.
Its motivation comes from the real Internet: Many autonomous systems are initially not available to new customers, but they may become as connectivity demand in their neighborhood grows.

Let $A \geq 1$ be an integer parameter. The FN algorithm maintains a set $F$ of fertile nodes; initially, only the first node is in $F$. Each new node $i > 0$ is connected to the fertile node optimizing the trade-off:

$$\min_{j \in F} (\alpha \ast d_{ij} + ecc(j)).$$

Finally, if the chosen node $j$ has more than $A$ infertile children, then the infertile child of $i$ that is closest to $j$ becomes fertile and is inserted to $F$.

In Figure 4 the value of $A$ is two. Thus, when node 0 has more than two infertile children, the closest one (node 1) becomes fertile.

4. Comparison of the Models

4.1. The Pandora System

There are many topology generators available. However, while implementing and analyzing the models in this paper, the authors realized that there was no single generator that could accommodate our models with reasonable effort. We therefore implemented our own network topology generator and analyzer, called Pandora, with these characteristics:

1. Large scale: Depending on the algorithm, the generator can produce graphs up to 80,000 nodes.
2. Extensible: Any researcher with a little knowledge on programming can write their own algorithms and metrics and add them to the generator. For example, once the user is accustomed to the features of the generator, the FKP algorithm can be implemented in 5 minutes.

3. Flexible: The generator can import files from inet, pandora and topgen formats and export data to otter and pajek formats.

4. User friendly: The simplicity of the design of the generator minimizes the time for a person to familiarize and use the generator. New users have started writing code within minutes.

5. Portable: Implemented in Java, the generator can run on any platform.

For more details, more experimental results, and for actually using and downloading Pandora, see [15].

4.2. The Metrics

We chose eight different metrics for evaluating and comparing our models; many of these metrics were proposed in [8].

1. The first three parameters are related to the degree distribution. The average degree (or, equivalently, the number of edges) is the most basic one.

2. The degree exponent $\gamma$, found by regression analysis on a log-log plot of the degree distribution. The relative frequency of nodes of degree $d$ is proportional to $k^{-\gamma}$.

3. The power law max degree is the natural cut-off of the degree probability density function, defined as $n^{\frac{1}{\gamma}-1}$, where $n$ is the number of nodes.

4. Another important aspect of graph structure is the joint degree distribution, looking at the degree distribution among the neighbors of a node with a given degree. Two summary statistics of the joint degree distribution are the average neighbor degree and the assortativity coefficient. The average neighbor degree of a $k-1$ degree node is equal to:

$$\sum_{k_2=1}^{D_{max}} k_2 \cdot P(k_2|k_1).$$

where $P(k_2|k_1)$ is a conditional probability that given a node with degree $k_1$ will be connected with a node of degree $k_2$ and $D_{max}$ is the maximum degree of the graph. Max neighbor degree is the average neighbor degree for $k_1 = D_{max}$. Average neighbor degree of the whole graph is the average of Eq.(5) for all the degrees.

5. The assortativity coefficient quantifies the connection between nodes of similar degree. If nodes of high degree tend to be connected with node of low degree then we say that the graph is disassortative, if they connect to nodes of similar degree the graph is called assortative. Notice that the assortativity coefficient and the average neighbor degree are very close in meaning: Graphs
with low average neighbor degree or is characterized as disassortative, and is vulnerable to random failures and virus attacks. On the other hand such graphs are better at traffic monitoring and prevention of DoS attacks [19].

6. We finally consider two key metrics related to clustering. Mean local clustering is the average value of $LC(k)$, that is the average number of links between neighbors of $k$-degree nodes.

7. Finally, the clustering exponent is the exponent of the log-log plot of local clustering vs. degree.

### 4.3. Results

Table 1 displays the results of calculating these metrics for the output of five of our models, and comparing them to the same metrics of the actual Internet graph obtained from [8]. We also compare our models with two classic models of internet topology, FKP and a variant of the classical BA model as produced by the inet 3.0 topology generator [22]. Each entry of the table represents the average of 15 runs. To compare the models effectively, the generated graphs have the same number $n = 5556$ of nodes as the Internet graph data in [8].

<table>
<thead>
<tr>
<th>BGP Tables</th>
<th>Reduced</th>
<th>FKP</th>
<th>BA</th>
<th>EBT</th>
<th>CD</th>
<th>IP</th>
<th>FN</th>
<th>EIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of edges</td>
<td>13473</td>
<td>5556</td>
<td>9876</td>
<td>12818</td>
<td>11095</td>
<td>11110</td>
<td>5556</td>
<td>12787</td>
</tr>
<tr>
<td>Average degree</td>
<td>4.85</td>
<td>2</td>
<td>3.55</td>
<td>4.61</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4.6</td>
</tr>
<tr>
<td>Frequent exponent ($\gamma$)</td>
<td>2.14</td>
<td>2.02</td>
<td>1.22</td>
<td>2.14</td>
<td>2.132</td>
<td>2.15</td>
<td>2.146</td>
<td>2.14</td>
</tr>
<tr>
<td>Power law max degree</td>
<td>1893</td>
<td>775</td>
<td>2577</td>
<td>1934</td>
<td>2011</td>
<td>1765</td>
<td>1888</td>
<td>1928</td>
</tr>
<tr>
<td>Assortativity coefficient</td>
<td>-0.21</td>
<td>-0.29</td>
<td>-0.17</td>
<td>-0.167</td>
<td>-0.21</td>
<td>-0.22</td>
<td>-0.23</td>
<td>-0.156</td>
</tr>
<tr>
<td>Norm. average neighbor degree</td>
<td>0.02</td>
<td>0.05</td>
<td>0.15</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.037</td>
<td>0.04</td>
</tr>
<tr>
<td>Norm. max neighbor degree</td>
<td>0.03</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.1</td>
<td>0.12</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Mean Local clustering</td>
<td>0.2</td>
<td>0</td>
<td>0.42</td>
<td>0.61</td>
<td>0.62</td>
<td>0.23</td>
<td>0</td>
<td>0.42</td>
</tr>
<tr>
<td>Clustering exponent</td>
<td>0.3</td>
<td>0</td>
<td>0.52</td>
<td>0.81</td>
<td>0.72</td>
<td>0.29</td>
<td>0</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Networks with high average degree tend to be better connected. As it was mentioned before, the enhanced best two and independent path models achieve an average degree of 4.6, which is very close to the real Internet graph data. Since the degree distribution is maybe the most commonly used topology characteristic, in all our models we adjusted parameters so as to agree with the Internet graph on this metric. All our models produce a very good frequent exponent and maximum power law degree.

As for the other metrics, the controlled distance, independent path and fertile node models tend to be much closer to the Internet data in terms of assortativity coefficient. Concerning neighbor degree metrics, we notice that most of our models are not close to the Internet with respect of the value of maximum neighbor degree but in terms of average neighbor degree achieve good results.

The fertile model is closer to the Internet graph data than the FKP model without fertility according to all metrics, and achieves a very good frequent exponent.
Finally, the independent path model is the most accurate with respect to the clustering metrics; the fertile model, always generating a tree, is, of course, highly inaccurate in this regard.

5. Conclusions and Future Work

The FKP model was a first attempt to take into account the complex and decentralized decision making involved in the evolution of the Internet; its power law behavior can be shown analytically, and its power law exponent is, empirically, quite accurate. On the negative side, the FKP model generates trees with no clustering and too many leaves. The extensions of the FKP model introduced and evaluated in this paper occupy a very interesting point in the accuracy-simplicity trade off, being quite simple and intuitive, while at the same time the producing graphs are rather Internet-like.

Furthermore, the experimental methodology (generic and flexible topology generator and standardized metric calculation) used in this paper may be useful in future work. The models proposed are simple enough that some of the crucial metrics may be possible to calculate analytically. We have some preliminary results on the exponent of the fertile node model [20], but much work is needed to be done on this front.

Finally, it would be interesting to incorporate in our models deletions of nodes and edges, certainly an element of the real Internet which, to our knowledge, has not been considered in the literature.

References


