Power-Laws and the AS-level Internet Topology

Georgos Siganos,Michalis Faloutsos,Petros Faloutsos,Christos Faloutsos

***Abstract*— In this paper, we study and characterize the topology of the Internet at the Autonomous System level. First, we show that the topology can be described efficiently with power-laws. The elegance and simplicity of the power-laws provide a novel perspective into the seemingly uncontrolled Internet structure. Second, we show that power-laws appear consistently over the last 5 years. We also observe that the power-laws hold even in the most recent and more complete topology [10] with correlation coefficient above 99% for the degree power-law. In addition, we study the evolution of the power-law exponents over the 5 year interval and observe a variation for the degree based power-law of less than 10%. Third, we provide relationships between the exponents and other topological metrics.**

1 Introduction

In this paper, we study the topology of the Internet and we identify several power-laws. Furthermore, we discuss multiple benefits from understanding the topology of the Internet. Our work is motivated by questions like the following *"What does the Internet look like?" "Are there any topological properties that don't change in time?" "How will it look like a year from now?" "How can I generate Internet-like graphs for my simulations?"*.

Modeling the Internet topology is an important open problem despite the attention it has attracted recently. Paxson and Floyd consider this problem as a major reason why we don't know how to simulate the Internet [21]. An accurate topological model can have significant impact on network research. First, we can design more efficient protocols that take advantage of its topological properties. Second, we can create more accurate artificial models for simulation purposes. And third, we can derive estimates for topological parameters (e.g. the average number of neighbors within h hops) that are useful for the analysis of protocols and for speculations of the Internet topology in the future.

In this paper, we propose the use of power-laws to describe the topology of the Internet at the Autonomous System or interdomain level. Power-laws are expressions of the form $y \propto x^a$, where a is a constant, x and y are the measures of interest, and $\propto$ stands for "proportional to". Conceptually, our work has three main thrusts: a) defining and identifying the power-laws, b) studying their evolution, and c) relating power-laws exponents and other graph metrics. Our work can be summarized in the following points.

First, we identify several power-laws that describe the distribution of topological metrics such as node degree. We also show that the three power-laws are tightly related theoretically. In addition, we introduce a graph metric to quantify the density of a graph and propose a power-law approximation of that metric.

Second, we study the evolution of the power-laws between November 1997 and February 2002. The power-laws hold for 1253 instances, with good linear fits in log-log plots; the correlation coefficient of the fit is at least 96%, typically above 98%. We note that their existence is persistent, and they hold even in the most recent and more complete topology [10].

Third, we present new and known relationships between power-laws exponents and other graph metrics. We list mechanisms that create power-laws, discuss their plausibility and their efficiency in creating graphs for practical purposes.

Our work in perspective. Power-laws is a first step in understanding the Internet topology. The evidence of their existence is too strong to be dismissed as coincidence. We monitor and analyze the Internet over a period of five years, during which the size of the network quadrupled. The contributing sources for the data collection changed significantly in number and location [41]. Additionally, we analyzed the more recent and complete topology [10]. These observations exclude by and large the possibility of the power-laws being the result of coincidence. Therefore, the power-laws appear as a necessary though not sufficient condition for a topology to be realistic. There may be more topological properties of the Internet topology that are not captured by our power-laws [45, 54].

The rest of this paper is structured as follows. In Section 2, we present some definitions and previous work on measurements and models for the Internet. In Section 3, we present our Internet instances and provide useful measurements. In Section 4, we present our three observed power-laws and our power-law approximation. In Section 5, we present the time evolution of the exponent of the power-laws we presented in the previous section. In Section 6, we present the models used to generate power-law graphs. In Section 7, we conclude our work and discuss future directions.

Georgos Siganos and Michalis Faloutsos are with the Deptartment of Computer Science, U.C. Riverside (email:{siganos;michalis}@cs.ucr.edu). Petros Faloutsos is with the Department of Computer Science, U.C. Los Angeles (email:pfal@cs.ucla.edu). Christos Faloutsos is with the Department of Computer Science, Carnegie Mellon University (email:christos@cs.cmu.edu). This research was supported by the Defense Advanced Research Projects Agency (DARPA) under grant N660001- 00-1-8936 and by the NSF Career ANIR 9985195.

2 Background and Previous Work

The Internet can be decomposed into subnetworks that are under separate administrative authorities. These subnetworks are called *domains* or *Autonomous Systems*. This way, the topology of the Internet can be studied at two different granularities. At the **router level**, we represent each router by a node [43]. At the **inter-domain level**, each domain is represented by a single node [22] and each edge is an inter-domain interconnection. The study of the topology at both levels is equally important. The Internet community develops and employs different protocols inside a domain and between domains. An intra-domain protocol is limited within a domain, while an inter-domain protocol runs between domains treating each domain as one entity. Here, we focus on the autonomous system level and represent the topology of the Internet by an undirected graph.

Symbol	Definition
G	An undirected graph.
N	Number of nodes in a graph.
E	Number of edges in a graph.
δ	The diameter of the graph.
d_v	Degree of node v.
d	Average degree of the nodes of a graph: $\bar{d} = 2\ E/N$

Table 1: Definitions and symbols.

Network analysis before power-laws. Before 1999, the metrics that were used to evaluate network models were mainly the node degree and the distances between nodes. Given a graph, the *degree* of a node is defined as the number of edges incident to the node (see Table 1). The distance between two nodes is the number of edges along the shortest path between the two nodes. Most studies report minimum, maximum, and average values and plot the degree and distance distribution. We denote the number of nodes of a graph by N, the number of edges by E, and the diameter of the graph by δ. Using these metrics Govindan and Reddy [22] study the growth of the inter-domain topology of the Internet between 1994 and 1995. The graph is sparse with 75% of the nodes having degrees less or equal to two. Pansiot and Grad [43] study the topology of the Internet in 1995 at the router level. The distances they report are approximately two times larger compared to those of Govindan and Reddy.

For graph generation purposes, Waxman introduced a popular network model [55]. The link creation probabilities depend upon the Euclidean distance between the nodes. This model was successful in representing small early networks such as the ARPANET. As the size and the complexity of the network increased more detailed models were needed [16] [8]. Zegura et al. [60] reviewed these generation methods using a more expansive set of metrics, including some that are driven by uses, like multicast routing. Based on the limitations they found, they introduced a comprehensive model that includes several previous models.

Power-laws: a ubiquitous presence. Pareto was among the first to introduce power-laws in 1896 [44]. He used power-laws to describe the distribution of income where there are few very rich people, but most of the people have a low income. Another classical law, the Zipf law [61], was introduced in 1949, for the frequencies of the English words and the population of cities. Power-laws have been found in numerous diverse fields spanning geological, natural, sociological, and biological systems. Some interesting examples of power-law distributions are the movie actor collaboration network [7], the human respiratory system [34], automobile networks [19], the size and location of earthquakes, stock-price fluctuations [6], the web of human sexual contacts [17], biological cellular networks [25], the scientific citation network [50]. More details about the historical aspects of power-laws can be found by Mitzenmacher [38] and an extensive presentation of power-laws in many diverse fields in Reka [3].

Network analysis using power-laws. More recently, power-laws have been observed in communication networks. First, power-laws have been observed in network traffic [56][30][46][13]. In addition, the topology of the World Wide Web [4, 28] can be described by power-laws. Furthermore, power-laws describe the topology of peer-to-peer networks [39] and properties of multicast trees [12, 47, 57, 37]. Among these properties, the Chuang-Sirbu law states that the size of the multicast tree follows a power-law with respect to the number of group members with exponent 0.8.

Our initial work [20] on power-laws has generated significant follow-up work. Various researchers verified our observations with different datasets[24, 23, 33]. In addition, significant work has been devoted in understanding the origin [36], and generating power-law topologies [35, 36, 42, 26, 54, 58]. We discuss these approaches for generating power-laws in section 6. More recently, several works have focused on describing the topology in a qualitatively way [53, 31, 32, 52].

3 Our Internet Instances

In this section, we present the Internet instances we study in our work. We use topologies from two sources. First, we use the Oregon routeviews project [41]. The information is collected by a route server from BGP[1][49] routing tables of multiple geographically distributed BGP routers. This is the *only* archival repository we could find in order to study the evolution of the topology. However, the Oregon data does not identify all possible links between ASs [10]. For this reason, we use a second data set from 2001 [10], which is the superset of Oregon and several other routing repositories. This data is currently considered as the most comprehensive AS topology although it is almost certainly not complete. Unfortunately, there is a limited number of these instances, which span only 9 weeks, starting from

[1]BGP stands for the Border Gateway Protocol, and is the inter-domain routing protocol.

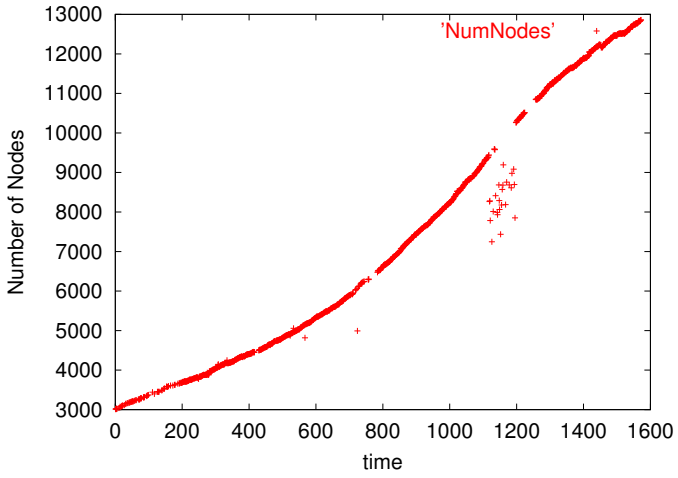


Figure 1: The growth of the Internet: the number of domains versus time between the end of 1997 until the start of 2002.

March 2001, and thus does not lend itself to an evolution study.

The Oregon dataset contains 1253 daily instances. These instances span an interval of 1600 days, more than five years, from 8th of November 1997 till 28th of February 2002. Note that we filter the data to remove incomplete data files that they do not represent correctly the topology. We identify and remove the instances that have less than 50% of the nodes found in the previous instance. For example, we removed the reported topology on the 29th of August 1999, which has 103 nodes, while the files on the previous and next day have more than 5600 nodes. Among the 1253 instances, we selected the instance of May 26th 2001 to demonstrate the power-laws, so that we can compare the results we have with the more complete topology. For the rest of this paper, we will refer to the instance from Oregon as **Oregon**, and the instance which represents the more complete topology as **Multi** respectively.

Note that the remaining 1252 instances, also follow the power-laws. Furthermore, the size of the topology in the five-year period quadrupled (see figure 1). The change is significant, and it ensures that our instances, reflect different snapshots of an evolving network.

4 Power-Laws of the Internet

In this section, we observe three power-laws of the Internet topology. We propose and measure graph properties, which demonstrate a regularity that is unlikely to be a coincidence. The exponents of the power-laws can be used to characterize graphs. In addition, we introduce a graph metric that is tailored to the needs of the complexity analysis of protocols. The metric reflects the density or the connectivity of nodes, and we offer a rough approximation of its value through a power-law. Finally, using our observations and metrics, we identify a number of interesting

Symbol	Definition
D_d	The Complementary Cumulative Distribution Function or CCDF, of a degree, is the percentage of nodes that have degree greater than the degree d.
r_v	The rank of a node, v, is its index in the order of decreasing degree.
$P(h)$	The *number of pairs* of nodes is the total number of pairs of nodes within less or equal to h hops, including self-pairs, and counting all other pairs twice.
NN(h)	The average number of nodes in a neighborhood of h hops.
λ	The eigenvalue of an NxN matrix A: $X : X \in \mathcal{R}^N \backslash \{0\}$ and $AX = \lambda X$.
i	The *order* of λ_i in the sorted sequence $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N$ of the eigenvalues of a matrix.

Table 2: Novel definitions and their symbols.

relationships between important graph parameters.

The goal of our work is to find metrics that quantify topological properties and describe concisely skewed data distributions. Previous metrics, such as the average degree, fail to do so. First, metrics that are based on minimum, maximum and average values are not good descriptors of skewed distributions; they miss a lot of information and probably the "interesting" part that we would want to capture. Second, the plots of the previous metrics are difficult to quantify, and this makes difficult the comparison of graphs.

To express our power-laws, we introduce several graph metrics that we show in Table 2. We define D_d to be the complementary cumulative distribution function of a degree, d, which is the percentage of nodes that have degree greater than the degree d. If we sort the nodes in decreasing degree sequence, we define *rank*, r_v, to be the index of the node in the sequence, while ties in sorting are broken arbitrarily. We define the number of pairs of nodes $P(h)$ to be the total number of pairs of nodes within less or equal to h hops, including self-pairs, and counting all other pairs twice. The use of this metric will become apparent later. We also define NN(h) to be the average number of nodes in a neighborhood of h hops. Finally, we recall the definition of the eigenvalues of a graph, which are the eigenvalues of its adjacency matrix.

We use linear regression to fit a line in a set of two-dimensional points [48]. The technique is based on the least-square errors method. The validity of the approximation is indicated by the correlation coefficient which is a number between -1.0 and 1.0. For the rest of this paper, we use the absolute value of the correlation coefficient, ACC. An ACC value of 1.0 indicates perfect linear corre-

lation, i.e., the data points are exactly on a line.

4.1 The rank exponent $\mathcal{R}$

In this section, we study the degrees of the nodes. We sort the nodes in decreasing order of degree, d_v, and plot the (r_v, d_v) pairs in log-log scale. The plots are shown in figure 2. The measured data is represented by points, while the solid line represents the least-squares approximation.

A striking observation is that the plots are approximated well by linear regression. The correlation coefficient is 0.97 for the Oregon, and 0.978 for the Multi topology. This leads us to the following power-law and definition.

Power-Law 1 (rank exponent) *Given a graph, the degree, d_v, of a node v, is proportional to the rank of the node, r_v, to the power of a constant, $\mathcal{R}$:*

$$d_v \propto r_v^{\mathcal{R}}$$

Definition 1 *Let us sort the nodes of a graph in decreasing order of degree. We define the rank exponent, $\mathcal{R}$, to be the slope of the plot of the degrees of the nodes versus the rank of the nodes in log-log scale.*

Intuitively, Power-Law 1 reflects a principle of the way domains connect; the linearity observed in 1253 graph instances is unlikely to be a coincidence.

Extended Discussion - Applications. We can estimate the proportionality constant for Power-Law 1, if we require that the minimum degree of the graph is m ($d_N = m$). This way, we can refine the power-law as follows.

Lemma 1 *In a graph where Power-Law 1 holds, the degree, d_v, of a node v, is a function of the rank of the node, r_v and the rank exponent, $\mathcal{R}$, as follows*

$$d_v = \frac{m}{N^{\mathcal{R}}} \; r_v^{\mathcal{R}}$$

Proof.

We can estimate the proportionality constant, C, for Power-Law 1, if we require that the degree of the N-th node is m, $d_N = m$.

$$\begin{aligned} d_N &= C\, N^{\mathcal{R}} \quad \Rightarrow \\ C &= m/N^{\mathcal{R}} \end{aligned} \tag{1}$$

We combine Power-Law 2 with Equation 1, and conclude the proof. ∎

Finally, using lemma 1, we relate the number of edges with the number of nodes and the rank exponent.

Lemma 2 *In a graph where Power-Law 1 holds, the number of edges, E, of a graph can be estimated as a function of the number of nodes, N, and the rank exponent, $\mathcal{R}$, as follows:*

$$E \approx \frac{1}{2\,(\mathcal{R}+1)} \; (1 - \frac{1}{N^{\mathcal{R}+1}}) \; N$$

Proof. The sum of all the degrees for all the ranks is equal to two times the number of edges, since we count each edge twice.

$$\begin{aligned} 2\,E &= \sum_{r_v=1}^{N} d_v \quad \Rightarrow \\ 2\,E &= \sum_{r_v=1}^{N} (r_v/N)^{\mathcal{R}} = (1/N)^{\mathcal{R}} \sum_{r_v=1}^{N} r_v^{\mathcal{R}} \quad \Rightarrow \\ E &\approx \frac{1}{2\,N^{\mathcal{R}}} \int_1^N r_v^{\mathcal{R}}\, dr_v \end{aligned} \tag{2}$$

In the last step, above we approximate the summation with an integral. Calculating the integral concludes the proof. ∎

Note that Lemma 2 can give us the number of edges as a function of the number of nodes for a given rank exponent. For an additional discussion on estimates using this formula see [20].

4.2 The Degree exponent $\mathcal{D}$

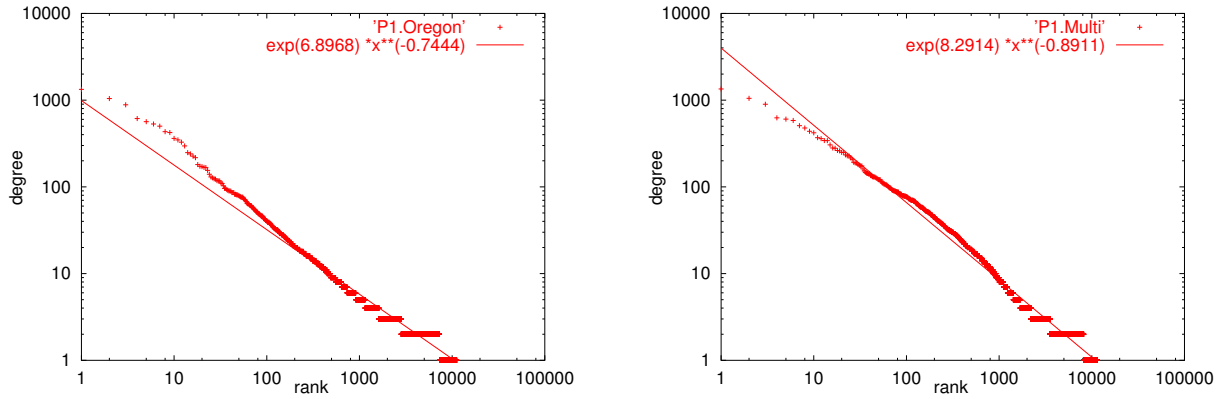
In this section, we study the distribution of the degree of the nodes. We plot the D_d versus the degree d in log-log scale in figure 3. The major observation is that the plots are approximately linear. The correlation coefficient is 0.996 for the Oregon and 0.991 for the Multi topology. As in the previous power-law, the slope of the exponent is different, something which is expected since the Multi topology has many more links. Note that in [10] they argue that this power-law doesn't hold for the Multi topology, without trying to approximate it using linear regression. Their conclusion is arguable, since we have a correlation coefficient of 0.991. Again as in the last power-law we checked to see if the power-law holds for all the instances we had. We found that the power-law holds for all the instances, and the correlation coefficient was always higher than 0.99. This leads us to the following power-law and definition.

Power-Law 2 (degree exponent)
Given a graph, the CCDF, D_d, of an degree, d, is proportional to the degree to the power of a constant, $\mathcal{D}$:

$$D_d \propto d^{\mathcal{D}}$$

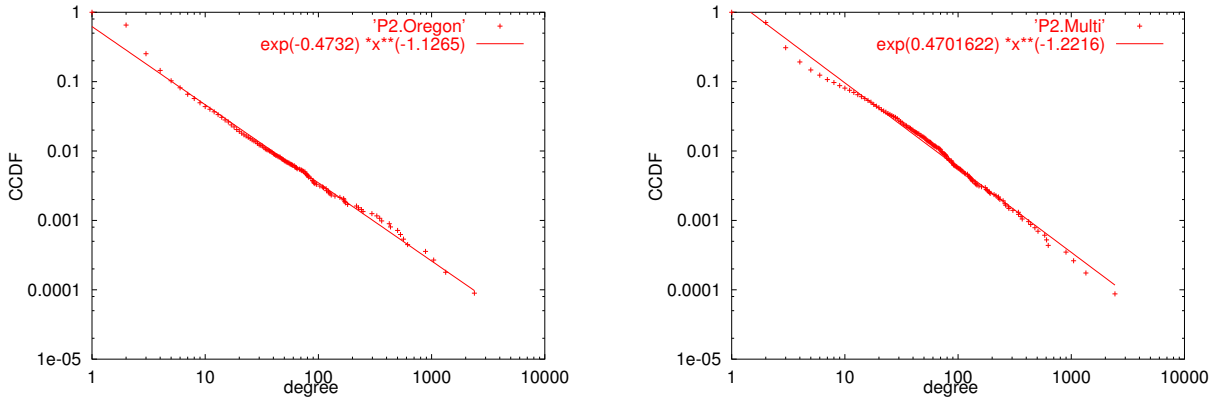
Definition 2 *We define the degree exponent, $\mathcal{D}$, to be the slope of the plot of the Cumulative degree of the degrees versus the degrees in log-log scale.*

The intuition behind this power-law is that the distribution of the degree of Internet nodes is not arbitrary. The qualitative observation is that degrees range over several orders of magnitude in a scale-invariant way. As a result, there is a non-trivial probability of finding nodes with very high degree. Our power-law manages to quantify this observation with the degree exponent. This way, we can test the realism of a graph with a simple numerical comparison. If a graph does not follow Power-Law 2, or if its degree exponent is considerably different from the real exponents, it probably does not represent a realistic topology.



(a) Oregon (b) Multi

Figure 2: The rank plot. Log-log plot of the degree d_v versus the rank r_v in the sequence of decreasing degree.



(a) Oregon (b) Multi

Figure 3: The Log-log plot of D_d versus the degree for the oregon and Multi topologies.

Comment. Note that the degree power-law that we present here is different than the one presented in our earlier work [20]. They both refer to the same distribution and their difference is that the previous power-law uses the probability distribution function[2], while the power-law here uses the cumulative distribution function. As a result, the exponents of the different power-laws differ approximately by one. Theoretically, the difference should be exactly one, since the cumulative distribution can be obtained by integrating the probability distribution. In practice, we see that the difference is not equal to one, due to approximations like the use of curve-fitting to find the slope. The cumulative distribution is preferable since it can be estimated in a statistically robust way.

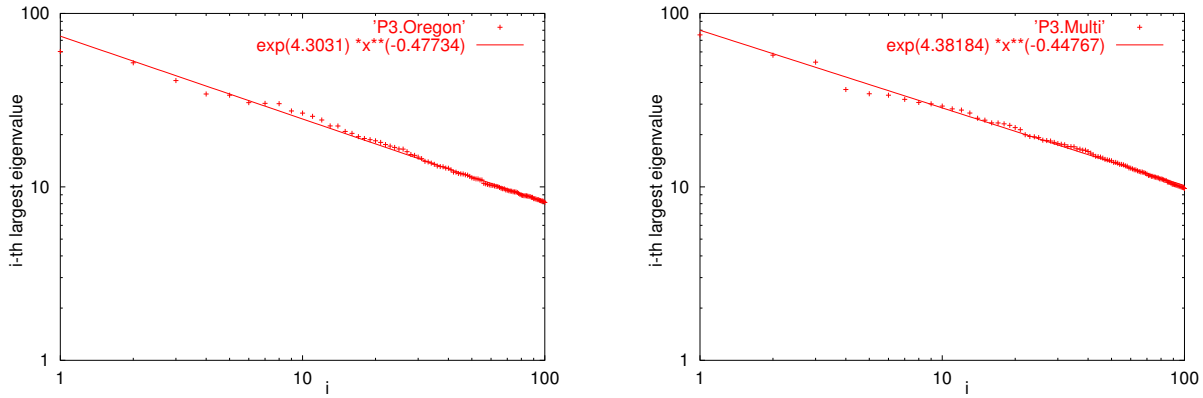
The relationship of the rank and degree power-laws. Both the rank and the degree power-laws characterize the degree distribution from different angles. It can be shown that the exponents of the two power-laws are related [61] [11][29]. More specifically, in a perfect power-law distribution, the slope of the rank power-law is equal to $\mathcal{R} = \frac{1}{\mathcal{D}}$. For the Multi topology the rank slope is 0.89. Using the above formula and the degree slope, we find the rank slope to be equal to 0.81. This discrepancy can be attributed to measurement imperfections and inaccuracies. In this regard, we think that it is useful to report both exponents when characterizing a topology.

[2]The actual law stated that: The frequency, f_d of a degree, d, is proportional to the degree to the power of a constant, $\mathcal{D}$, where the frequency, f_d, of a degree is the number of nodes with degree d.

4.3 The eigen exponent $\mathcal{E}$

In this section, we identify properties of the eigenvalues of our Internet graphs. Recall that the eigenvalues of a graph are the eigenvalues of its adjacency matrix. We plot the eigenvalue λ_i versus i in log-log scale for the first 100 eigenvalues. Recall that i is the order of λ_i in the decreasing sequence of eigenvalues. The result is shown in figure 4. The eigenvalues are shown as points in the figures, and the solid lines are approximations using a least-squares fit. Similar observations with equally high correlation coefficients were observed for all the other instances. We observe that the plots are practically linear with a correlation coefficient of 0.996 for both plots. The eigen exponent is -0.477 for the Oregon and -0.447 for the Multi topology.

It is rather unlikely that such a canonical form of the



(a) Oregon (b) Multi

Figure 4: The plot of the hundred largest eigenvalues for the Oregon and Multi topologies.

eigenvalues is purely coincidental, and we therefore conjecture that it constitutes an empirical power-law of the Internet topology.

Power-Law 3 (eigen exponent) *Given a graph, the eigenvalues, λ_i, are proportional to the order, i, to the power of a constant, $\mathcal{E}$:*

$$\lambda_i \propto i^{\mathcal{E}}$$

Definition 3 *We define the eigen exponent, $\mathcal{E}$, to be the slope of the plot of the sorted eigenvalues versus their order in log-log scale.*

Eigenvalues are fundamental graph metrics. There is a rich literature that proves that the eigenvalues of a graph are closely related to many basic topological properties such as the diameter, the number of edges, the number of spanning trees, the number of connected components, and the number of walks of a certain length between vertices, as we can see in [14]. All of the above suggest that the eigenvalues intimately relate to topological properties of graphs. However, it is not trivial to explore the nature and the implications of this power-law.

The relationship of the degree and the eigenvalue power-laws. A surprising relationship exists between the two exponents: the eigenvalue exponent is approximately the half of the degree exponent. In more detail, Mihail et al. [40] show that if the degrees $d_1, ..., d_n$ of graph follow a power-law, then the non-increasing sequence of the largest eigenvalues λ_i has the following one to one correspondence: $\lambda_i = \sqrt{d_i}$. It is worth noting that this is an asymptotic limit of the eigenvalues. If we take the logarithm of the previous equation, it follows that the two exponents differ by a factor of two. In practice, the exponents obey adequately the mathematical relationship, although the match is naturally not perfect. For example, the degree exponent for the Oregon topology is 1.12, and the eigenvalue exponent 0.47 which yields a ratio of 0.52 instead of 0.5.

4.4 The hop-plot exponent $\mathcal{H}$

In this section, we quantify the connectivity and distances between the Internet nodes in a novel way. We choose to study the size of the neighborhood within some distance, instead of the distance itself. Namely, we use the total number of pairs of nodes $P(h)$ within h hops, which we define as the total number of pairs of nodes within less or equal to h hops, including self-pairs, and counting all other pairs twice.

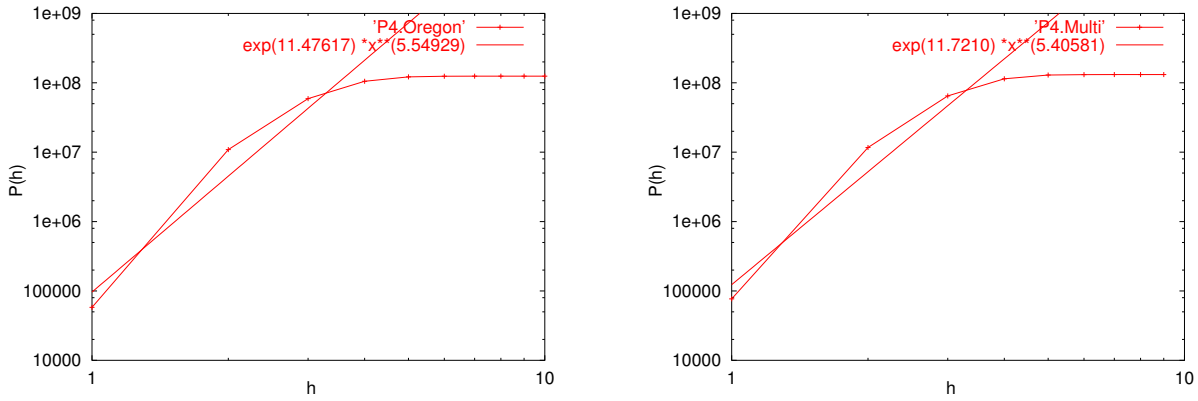
In figure 5, we plot the number of pairs $P(h)$ as a function of the number of hops h in log-log scale. The data is represented by points. We want to describe the plot by a line in least-squares fit, for $h \ll \delta$, shown as a solid line in the plots. We approximate the first 4 hops in the inter-domain graphs. The correlation coefficient is 0.9765 and 0.9784 for the Oregon and Multi topology respectively.

Unfortunately, four points is a rather small number to verify or disprove a linearity hypothesis experimentally. However, even this rough approximation has several useful applications as we show later in this section. It is worth mentioning that Philips et al. [47] state that the neighborhood growth is exponential and not a power-law. In figure 6, we plot again the number of pairs in log-lin for the Multi topology. We approximate the first four hops and found a correlation coefficient of 0.918 which is much lower than the previous correlation. From this, it seems that we can approximate the hopplot better with a power-law than with an exponential function.

Approximation 1 (hop-plot exponent) *The total number of pairs of nodes, $P(h)$, within h hops, is proportional to the number of hops to the power of a constant, $\mathcal{H}$:*

$$P(h) \propto h^{\mathcal{H}}, \quad h \ll \delta$$

Definition 4 *Let us plot the number of pairs of nodes, $P(h)$, within h hops versus the number of hops in log-log scale. For $h \ll \delta$, we define the slope of this plot to be the* hop-plot exponent, *$\mathcal{H}$.*



(a) Oregon (b) Multi

Figure 5: The hop-plot: Log-log plots of the number of pairs of nodes $P(h)$ within h hops versus the number of hops h.

Extended Discussion - Applications. We can refine Approximation 1 by calculating its proportionality constant. Let us recall the definition of the number of pairs, $P(h)$. For $h = 1$, we consider each edge twice and we have the self-pairs, therefore: $P(1) = N + 2\ E$. We demand that Approximation 1 satisfies the previous equation as an initial condition.

Lemma 3 *The number of pairs within h hops is*

$$P(h) \quad = \quad \begin{cases} c\ h^{\mathcal{H}}, & h \ll \delta \\ N^2, & h \geq \delta \end{cases}$$

where $c = N + 2\ E$ to satisfy initial conditions.

In networks, we often need to reach a target without knowing its exact position [51] [9]. In these cases, selecting the extent of our broadcast or search is an issue[3]. On the one hand, a small broadcast may not reach our target. On the other hand, an extended broadcast creates too many messages and takes a long time to complete. Ideally, we want to know how many hops are required to reach a "sufficiently large" part of the network. In our hop-plots, a promising solution is the intersection of the two asymptote lines: the horizontal one at level N^2 and the asymptote with slope $\mathcal{H}$. We calculate the intersection point using Lemma 3, and we define:

Definition 5 (effective diameter) *Given a graph with N nodes, E edges, and $\mathcal{H}$ hop-plot exponent, we define the effective diameter, δ_{ef}, as:*

$$\delta_{ef} = \left(\frac{N^2}{N + 2\ E} \right)^{1/\mathcal{H}}$$

Intuitively, the effective diameter can be understood as follows: any two nodes are within δ_{ef} hops from each other with high probability. We verified the above statement experimentally. The effective diameters of our inter-domain graphs was slightly over four. Rounding the effective diameter to four, approximately 80% of the pairs of nodes are within this distance. The ceiling of the effective diameter is five, which covers more than 95% of the pairs of nodes. The above confirms that the effective diameter manages to capture the majority of the distances. Furthermore, it argues indirectly that the hopplot exponent as a metric seems useful.

An advantage of the effective diameter is that it can be calculated easily, when we know N, and $\mathcal{H}$. Recall that we can calculate the number of edges from Lemma 2. Therefore, given estimates the hop-plot and rank-plot exponents, we can calculate the effective diameter of future Internet instances of a given size [20].

Furthermore, we can estimate the average size of the neighborhood, $\mathrm{NN}(h)$, within h hops using the number of pairs $P(h)$. Recall that $P(h) - N$ is the number of pairs without the self-pairs.

$$\mathrm{NN}(h) = \frac{P(h)}{N} - 1 \tag{3}$$

Using Equation 3 and Lemma 3, we can estimate the average neighborhood size.

Lemma 4 *The average size of the neighborhood, $NN(h)$, within h hops as a function of the hop-plot exponent, $\mathcal{H}$, for $h \ll \delta$, is*

$$NN(h) = \frac{c}{N}\ h^{\mathcal{H}} - 1, h > 0$$

where $c = N + 2\ E$ to satisfy initial conditions.

The average neighborhood is a commonly used parameter in the performance of network protocols. Our estimate is an improvement over the commonly used estimate that uses the average degree [59] [51] which we call **average-degree estimate**:

$$\mathrm{NN}'(h) = \bar{d}\ (\bar{d} - 1)^{h-1}$$

In figure 7, we plot the actual and the two estimates of the average neighborhood size versus the number of

[3]This problem has direct practical importance. The Internet has a built in mechanism for limiting the number of hops a packets makes. The time-to-live field of a packet is a counter that is decreased at each hop until it reaches zero, at which point the packet is not forwarded further.

hops using an instance from 1998. The superiority of the hop-plot exponent estimate is apparent compared to the average-degree estimate. The discrepancy of the average-degree estimate can be explained if we consider that the estimate does not comply with the real data; it implicitly assumes that the degree distribution is uniform. In more detail, it assumes that each node in the periphery of the neighborhood adds $\bar{d} - 1$ new nodes at the next hop. Our data shows that the degree distribution is highly skewed, which explains why the use of the hop-plot estimate gives a better approximation.

The most interesting difference between the two estimates is qualitative. The average degree based estimate considers the neighborhood size exponential in the number of hops. Our estimate considers the neighborhood as an $\mathcal{H}$-dimensional sphere with radius equal to the number of hops, which is a novel way to look at the topology of a network. Our data suggests that the hop-plot exponent-based estimate gives a closer approximation compared to the average-degree-based metric.

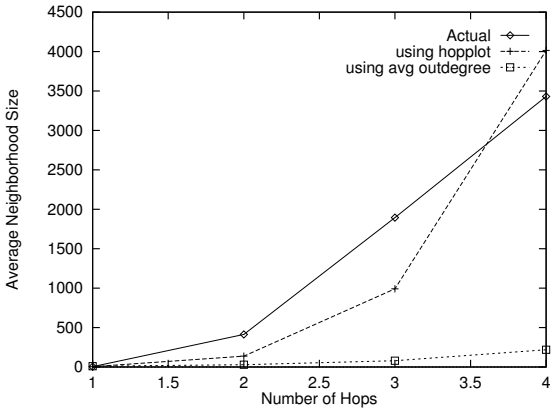


Figure 7: Average neighborhood size versus number of hops the actual, and estimated size a) using hop-plot exponent, b) using the average degree

5 The Persistence of Power-Law Exponents

We examine the evolution of power-law exponents in the five year span from November 1997 till February 2002. We want to stress that the main observation is that the power-laws hold for every instance. The evolution of the slope is a secondary issue.

The evolution of rank exponent $\mathcal{R}$. In figure 8, we examine the time evolution of the slope of the rank exponent. We plot the rank exponent versus the day that the instance of the graph was collected. The rank exponent $\mathcal{R}$ power-

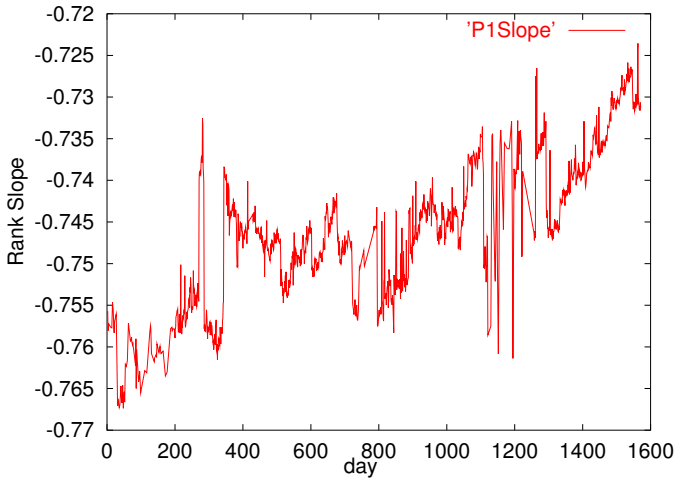


Figure 8: The evolution of the slope of the rank exponent

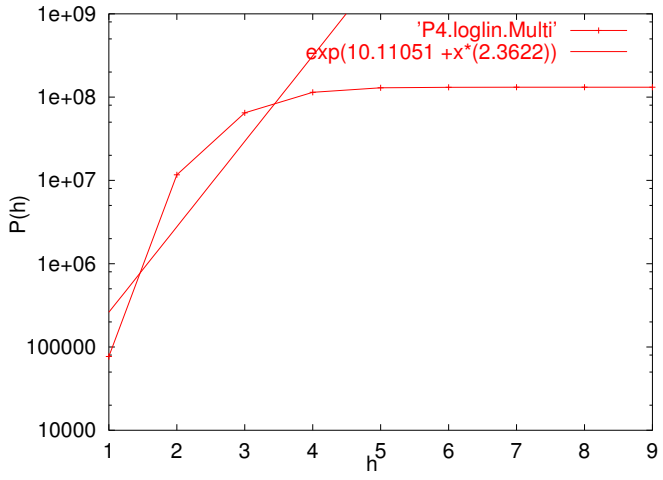


Figure 6: Approximating the hop-plot with an exponential function. This is a linear-logarithmic plot of the number of pairs of nodes $P(h)$ within h hops versus the number of hops h.

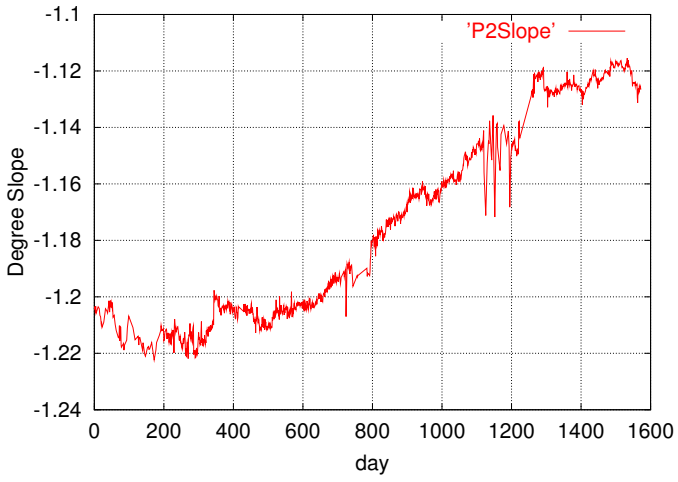


Figure 9: The evolution of the slope of the degree exponent

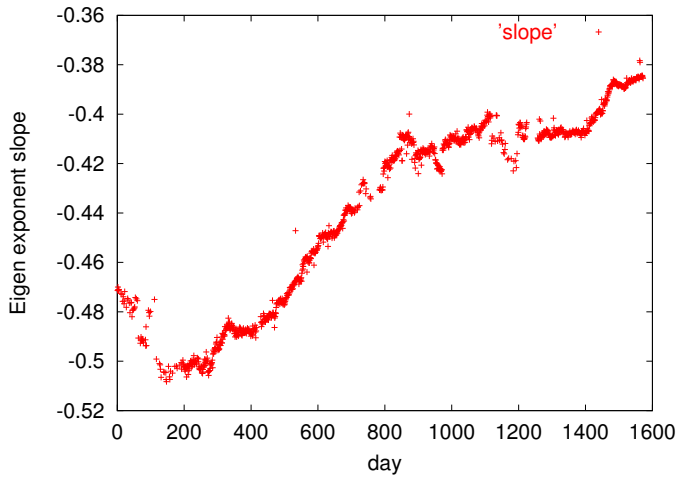


Figure 10: The evolution of the slope of the eigen exponent

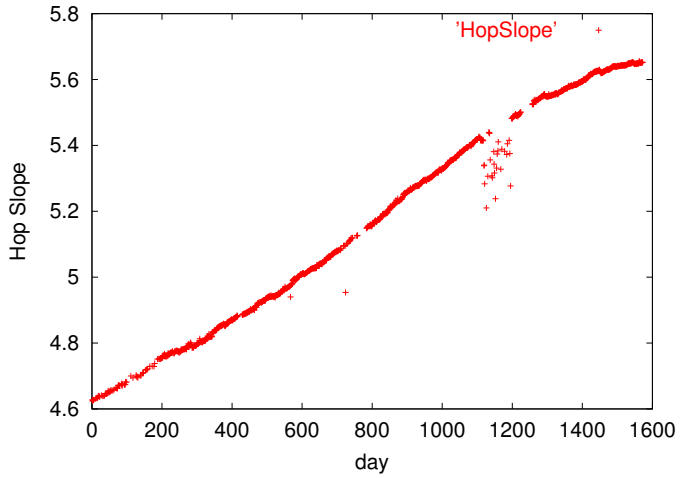


Figure 11: The evolution of the slope of the hop-plot exponent

law holds for all the instances, over a period of five years. The correlation coefficient of the law seems to decrease over time and for the last instance it is close to 0.9654. This indicates that the rank exponent, for the topologies from Oregon, should be treated with care in the future. However, we see in figure 2, that for the more complete topology we have a higher correlation coefficient.

The evolution of degree slope $\mathcal{D}$. We study the slope of the cumulative degree exponent and its evolution in time. In figure 9, we plot the degree exponent versus time. The degree exponent power-law holds for all the instances with a correlation coefficient always higher than 0.99. We observe from the graph that the slope is between -1.12 and -1.22, i.e. a variation of less than 9%.

The evolution of eigen exponent $\mathcal{E}$. In figure 10, we plot the time evolution of the eigen exponent. The power-law holds for all the instances we have measured. As we can see from the graph the value of the eigen exponent decreases for the first 150 instances and then it starts to rise again for the rest of the instances. We do not have an intuitive explanation for this behavior. Note that the eigenvalues of a graph does not depend on the way the nodes are enumerated.

hops	11-08-1997	02-28-2002
1	0.08%	0.04%
2	8.86%	8.09%
3	43.40%	46.64%
4	80.99%	84.76%
5	96.70%	97.46%
6	99.65%	99.71%

Table 3: The size of the neighborhood of a node (as percentage of the total) as a function of the hops (radius of neighborhood).

The evolution of hop-plot exponent. In figure 11, we plot the time evolution of the hop-plot exponent's slope. The power-law holds for all the instances with a correlation coefficient always higher than 0.97. We observe that the value of slope increases steadily. The initial value of the hop-plot exponent is 4.6 and for the latest instance is 5.7.

Understanding the hop-plot increase. As we saw, the network size increases significantly, while the distances between nodes increase very little. In table 3, we list the percentage of nodes that we can reach as a function of the number of hops, or the neighborhood of a node within h hops. We compare two graph instances, the 8th November of 1997 and our last instance the 28th of February 2002. Although the size of the graph quadrupled, we reach approximately the same percentage of nodes with the same number of hops. In absolute numbers, the number of nodes we can reach in 6 hops increased from approximately 3000 to 13000.

6 The Generation Of Power-laws

Why would such an uncontrolled[4] entity like the Internet follow any statistical regularities? Note that the high correlation coefficients rule out the possibility of pure coincidence. Intrigued by the previous question, and by the appearance of power-laws in many diverse fields, many scientists have tried to find the mechanism responsible for the creation of power-law graphs.

In this section, we will first give a small review of the most popular models that try to explain the appearance of power-laws in networks. Later we will briefly describe the status on the graph generation tools.

Scale-free networks. A very elegant growth model has been proposed by Barabasi and Reka. In their original

[4]The term uncontrolled refers to the fact that the Internet is not governed by a central authority, and it's growth and design is driven by many different optimization goals, such as financial, business and performance related.

work [7], their model states that the scale-free nature roots in two mechanism, the addition of new nodes, and their preferential attachment. Their model grows a graph by adding nodes. The probability of a new node connecting with node i of degree d_i is proportional to its degree: $\frac{d_i}{\sum d_j}$, where $\sum d_j$ is the sum of the degrees of all current nodes. In a more recent work, the same authors propose a more general model that includes generation of edges between existing nodes and rewiring [2], that is the removal of one edge and the creation of another between existing nodes. An extensive review on variations of the original idea that include more parameters can be found in [3].

There exists a number of real data studies based on the growth model proposed by Barabasi. In [10], they conclude that this theoretical model is not supported by the real data. Instead they mention that rewiring occurs infrequently, and that new nodes express a greater preference for nodes with large degree than is represented by the simple linear preference model. On the other hand, in [3, 58] and in [5] the authors show that the addition of nodes and edges follows the linear preferential model. This is controversial and more work is needed to compare the different approaches used.

Heuristically Optimized Trade-offs In [18], Fabrikant et al. propose "a simple and primitive model of Internet growth". In their model the power-law distributions root from the Internet growth in which two objectives are optimized simultaneously. The connection costs (last mile), and transmission delays measured in hops. Their model works as follows, they use a unit square plane, where nodes arrive and their place is chosen uniformly at random. Each node attaches itself on one of the previous nodes. They use two metrics in order to choose where the node should attach. The first metric is the Euclidean distance d_{ij} between the new node i and a node j. This metric captures the "last mile" costs. The second metric is a measure of the centrality of a node j h_j, if the new node attached to node j. This shows how close is the node to the center, and they mention that this captures the operation costs due to communication delays. Node i chooses to connect with node j that minimizes the weighted sum $min_{j<i} a d_{ij} + h_j$, where a is used to change the relative importance of the two objectives.

Highly Optimized Tolerance. In [27], Carlson et al. propose that power-laws are the result of an optimization, either through natural selection or engineering design, to provide robust performance despite uncertain environments. Regarding the Internet they mention that the survivability built in the Internet and its protocols can be the cause of the power-laws.

Topology Generators. The introduction of power-laws [20] brought a revision of the graph generation models in the networking community. The power-laws can be used as one question in the "qualifying exam" for the realism of a graph. The early generators failed when tested against power-laws, so after that a number of new generators was proposed [35, 36, 42, 26, 54, 58].

There are two kinds of graph generation tools. In the first one we have the tools that take the power-laws as given and they don't attempt to emulate the process that leads to a power-law [26, 1, 42]. In the second category, [35, 36, 54, 58] we have generators which try to capture the actual process that governs the creation of power-laws. All of them use variations of the preferential attachment model, described in [7, 2].

In the most recent effort [54], Bu et al. proposed a new generator that generates more realistic Internet topologies. Furthermore, they use additional metrics found in small world networks [15]. They show that previous generators fail in some of these new criteria. They show that by deviating from the linear preferential model by giving higher preferentiality to high degree nodes, they generate more realistic topologies.

7 Conclusions

In this paper, we propose power-laws as a tool to describe the Internet topology and examine their persistence in time. The power-laws capture concisely the highly skewed distributions of the graph properties. Finally, we show how these exponents relate to each other and how they relate to other topological properties.

We note the persistence of power-laws in time: they appear in more than 1200 daily instances over the span of more than five years from 1997 till 2002. In this interval, the network underwent significant changes in size (400%) and rate of growth. The monitoring infrastructure changed and evolved as well. This suggests that the appearance of power-laws is unlikely to be a coincidence or an artifact. An orthogonal but also striking observation is that some of the exponents did not change more than 10%. Furthermore, the power-laws seem to hold even in the most complete topology, which combines multiple sources. In fact, some of the power-laws hold with higher correlation coefficients in this data set.

Some additional observations can be summarized in the following points:

- Power-law exponents are a more efficient way to describe the highly-skewed graph metrics compared to average values of real graphs.
- We propose the number of pairs, $P(h)$, within h hops, as a metric of the density of the graph and approximate it using the hop-plot exponent, $\mathcal{H}$.
- We derive formulas that link the exponents of our power-laws with graph metrics such as the number of nodes, the number of edges, and the average neighborhood size.
- Using power-laws we obtain better intuition, for example we can see that the network becomes denser by observing the hop-plot exponent.

Apart from their theoretical interest, our power-laws have practical applications. First, our power-laws can assess the realism of synthetic graphs, and enhance the validity of simulations. Second, they can help analyze the average-

case behavior of network protocols. For example, we can estimate the message complexity of protocols using our estimate for the neighborhood size. Third, the power-laws can help answer "what-if" scenarios like *"what will be the diameter of the Internet, when the number of nodes doubles?" "what will be the number of edges then?"*

FUTURE WORK. The topological power-laws presented here form a critical step towards understanding and modeling the Internet. However, there are several open questions. First, we would like to explore further the meaning and the value of the exponents. We believe that such analysis could reveal interesting inter-plays and trade-offs between the forces that govern the creation of the topology. Second, the power-laws alone may not be sufficient in describing the topology in all its complexity. For example, we would like to develop more structural properties that will quantify the topology in a way that is easier to visualize. The goal of this direction is to develop a simple and intuitive model for the Internet topology.

ACKNOWLEDGMENTS. We would like to thank Mark Craven, Daniel Zappala, and Adrian Perrig for their help in earlier phases of this work. We would also like to thank Vern Paxson, and Ellen Zegura for their valuable feedback.

References

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *STOC*, 2000.

[2] R. Albert and A. Barabasi. Topology of complex networks:local events and universality. *Physical Review Letters*, 85,5234, 2000.

[3] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, 74, 47, 2002.

[4] R. Albert, H.Jeong, and A.L. Barabasi. Diameter of the world wide web. *Nature*, 401, 1999.

[5] A.Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of Internet. *Phys. Rev. E*, 65, 2002.

[6] P. Bak. *How Nature Works: The Science of Self Organized Criticality*. Springer-Verlag, 1996.

[7] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 8, October 1999.

[8] Kenneth L. Calvert, Matthew B. Doar, and Ellen W. Zegura. Modeling internet topology. *IEEE Communications Magazine*, 35(6):160–163, June 1997.

[9] K. Carlberg and J. Crowcroft. Building shared trees using a one-to-many joining mechanism. *ACM Computer Communication Review*, pages 5–11, January 1997.

[10] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger. The origin of power laws in Internet topologies revisited. *infocom*, 2002.

[11] H. Chou. A note on power-laws of Internet topology. *e-print cs.NI/0012019*, 2000.

[12] J. Chuang and M. Sirbu. Pricing multicast communications: A cost based approach. *In Proc. of the INET'98*, 1998.

[13] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic, evidence and possible causes. *SIGMETRICS*, pages 160–169, 1996.

[14] D. M. Cvetkovič, M. Boob, and H. Sachs. *Spectra of Graphs*. Academic press, 1979.

[15] D.J.Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[16] M. Doar. A better model for generating test networks. *Proc. Global Internet, IEEE*, Nov. 1996.

[17] Liljeros F., C. Edling, L.A.N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature 411*, 2001.

[18] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the Internet. *Extended Abstract STOC*, 2002.

[19] Christos Faloutsos and Ibrahim Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 24-26 1994. Also available as CS-TR-3198, UMIACS-TR-93-130.

[20] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *ACM SIGCOMM*, pages 251–262, Sep 1-3, Cambridge MA, 1999.

[21] S. Floyd and V. Paxson. Difficulties in simulating the Internet. *IEEE/ACM Transactions on Networking*, 2001.

[22] R. Govindan and A. Reddy. An analysis of Internet Inter-domain topology and route stability. *Proc. IEEE INFOCOM*, Kobe, Japan, April 7-11 1997.

[23] R. Govindan and H. Tangmunarunkit. Heuristics for Internet map discovery. *Proc. IEEE INFOCOM*, Tel Aviv, Israel, March 2000.

[24] S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang. On the placement of Internet instrumentation. *Proc. IEEE INFOCOM*, Tel Aviv, Israel, March 2000.

[25] H. Jeong, B. Tomber, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature 407 651*, 2000.

[26] Cheng Jin, Qian Chen, and Sugih Jamin. Inet: Internet topology generator. *Techical Report UM CSE-TR-433-00*, 2000.

[27] J.M.Carlson and J.Doyle. Highly optimized tolerance: a mechanism for power laws in designed systems. *Physics Review E*, 60(2), 1999.

[28] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A.Tomkins, and E. Upfal. The web as a graph. *ACM Symposium on Principles of Database Systems*, 2000.

[29] L.A.Adamic. Zipf, power-laws, and pareto - a ranking tutorial. *http://www.parc.xerox.com/iea/*, 2000.

[30] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic. *IEEE Transactions on Networking*, 2(1):1–15, February 1994. (earlier version in SIGCOMM '93, pp 183-193).

[31] L.Gao. On inferring automonous system relationships in the internet. *IEEE/ACM Transactions on Networking*, 9(6), 2001.

[32] L.Subramanian, S.Agarwal, J.Rexford, and R.Katz. Characterizing the Internet hierarchy from multiple vantage points. *Proc. IEEE INFOCOM*, 2002.

[33] Damien Magoni and Jean Jacques Pansiot. Analysis of the autonomous system network topology. *ACM Computer Communication Review*, July 2001.

[34] B. Mandelbrot. *Fractal Geometry of Nature*. W.H. Freeman, New York, 1977.

[35] A. Medina, A. Lakhina, I. Matta, and J. Byers. Brite:an approach to universal topology generation. *MASCOTS*, 2001.

[36] A. Medina, I. Matta, and J. Byers. On the origin of powerlaws in Internet topologies. *ACM SIGCOMM Computer Communication Review*, 30(2):18–34, April 2000.

[37] P. Van Mieghem, G. Hooghiemstra, and R. van der Hofstad. On the efficiency of multicast. *IEEE/ACM Transactions on Networking*, 9, 2001.

[38] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Allerton*, 2001.

[39] M.Jovanovic. Modeling large-scale peer-to-peer networks and a case study of gnutella. *Master thesis,University of Cincinnati*, 2001.

[40] M.Mihail and C.H.Papadimitriou. On the eigenvalue power law. *Random*, 2002.

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

[42] Christopher R. Palmer and J. Gregory Steffan. Generating network topologies that obey power laws. *IEEE Globecom*, 2000.

[43] J.-J. Pansiot and D Grad. On routes and multicast trees in the Internet. *ACM SIGCOMM Computer Communication Review*, 28(1):41–50, January 1998.

[44] V. Pareto. Cours d'economic politique. *Dronz,Geneva Switzerland*, 1896.

[45] K. Park and H. Lee. On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law Internets. *ACM SIGCOMM*, San Diego, Aug 2001.

[46] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995. (earlier version in SIGCOMM'94, pp. 257-268).

[47] G. Philips, S. Shenker, and H. Tangmunarunkit. Scaling of mul-

ticast trees: Comments on the chuang-sirbu scaling law. *ACM SIGCOMM*, Sep 1999.

[48] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

[49] Y. Rekhter and T. Li (Eds). A Border Gateway Protocol 4 (BGP-4). RFC 1771, 1995.

[50] S.Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. Jour. B 4, 131-134*, 1998.

[51] S.Yan, M.Faloutsos, and A.Banerjea. Qos-aware multicast routing for the Internet: The design and evaluation of Qosmic. *IEEE/ACM Transactions on Networking*, 10:54–66, February 2002.

[52] Hongsuda Tangmunarunkit, Ramesh Govindan, Sugih Jamin, Scott Shenker, and Walter Willinger. Network topology generators: Degree-based vs structural. *ACM SIGCOMM*, 2002.

[53] L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the Internet topology. *Global Internet, San Antonio, Texas*, 2001.

[54] T.Bu and D. Towsley. On distinguishing between Internet power law topology generators. *Proc. IEEE INFOCOM*, 2002.

[55] B. M. Waxman. Routing of multipoint connections. *IEEE Journal of Selected Areas in Communications*, pages 1617–1622, 1988.

[56] Walter Willinger, Murad Taqqu, Robert Sherman, and Daniel V. Wilson. Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level. *ACM SIGCOMM'95. Computer Communication Review*, 25:100–113, 1995.

[57] T. Wong and R. Katz. An analysis of multicast forwarding state scalability. *International Conference on Network Protocols*, 2000.

[58] S.H. Yook, H.Jeong, and A. Barabasi. Modeling the Internet's large-scale topology. *submitted for publication*, 2001.

[59] D. Zappala, D. Estrin, and S. Shenker. Alternate path routing and pinning for interdomain multicast routing. Technical Report USC CS TR 97-655, U. of South California, 1997.

[60] E. W. Zegura, K. L. Calvert, and M. J. Donahoo. A quantitative comparison of graph-based models for internetworks. *IEEE/ACM Transactions on Networking*, 5(6):770–783, December 1997. http://www.cc.gatech.edu/projects/gtitm/.

[61] G.K. Zipf. *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*. Addison Wesley, Cambridge, Massachusetts, 1949.