This document is downloaded from the
Digital Open Access Repository of VTT

Title	On the effect of very large nodes in Internet graphs
Author(s)	Reittu, Hannu; Norros, Ilkka
Citation	IEEE Global Telecommunications Conference, Globecom'02. Taipei, Taiwan, 17-21 Nov. 2002. Vol. 3, pp. 2624-2628
Date	2002
URL	http://dx.doi.org/10.1109/GLOCOM.2002.1189105
Rights	Copyright © (2002) IEEE. Reprinted from IEEE Global Telecommunications Conference, Globecom'02. Taipei, Taiwan. This article may be downloaded for personal use only

VTT
http://www.vtt.fi
P.O. box 1000
FI-02044 VTT
Finland

By using VTT Digital Open Access Repository you are bound by the following Terms & Conditions.

I have read and I understand the following statement:

This document is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of this document is not permitted, except duplication for research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered for sale.

On the effect of very large nodes in Internet graphs

Hannu Reittu and Ilkka Norros
VTT Information Technology
P.O. Box 1202, FIN-02044 VTT, Finland
{Hannu.Reittu, Ilkka.Norros}@vtt.fi

***Abstract*— We analyse a random graph where the node degrees are (almost) independent and have a distribution with finite mean but infinite variance — a region observed in empirical studies of the Internet. We show that the existence of very large nodes has a great influence on the connectivity. If N denotes the number of nodes, it seems that the distance between two randomly chosen nodes of the giant component grows as slowly as $\log\log(N)$. The essential observation is that very large nodes form a spontaneously arising "core network", which plays a crucial role in the connectivity, although its proportional size goes to zero as $N \to \infty$. Several results related to the core are proven rigorously, and a sketch of a full proof is given. Some simulations providing illustration of the findings are presented. Consequences of the results are discussed.**

I. Introduction

When the whole Internet is considered as a huge graph, where the nodes (vertices) are the routers and the edges are the links connecting them, it has been found that certain important characteristics follow power laws [1]. Moreover, a qualitatively similar picture arises when the nodes are Internet domains instead of individual routers. These observations may turn out to be significant for topology and routing dependent features of the Internet. This may even be compared with the role that the power laws of data traffic in communication networks play in modern queuing theory. Random networks with power law distribution of degrees of the nodes have been studied quite extensively in the past few years [2], [3], [4], [5], [6], [7]. They have several interesting practical applications, one of which is the topology of Internet.

In this paper, we adopt an analytically tractable random graph model, introduced by Newman, Strogatz and Watts [5]. The node degrees are independent and identically distributed random variables (except for one node with degree at most one). The probability mass function of node degree D behaves as $\mathbb{P}(D = d) \approx d^{-\tau}$, where $\tau > 1$. According to [1], the Internet graphs seem to fall in the region $\tau \in (2,3)$, which means that D has a finite mean but an infinite variance.

In the regime $\tau > 3$ it has been shown that the expectation of the diameter of the graph in number of hops scales as $\log(N)$, where N denotes the number of nodes [5]. Such graphs are sometimes called "small worlds". The method used in [5] does not work for $\tau < 3$. The authors suggest an exponential cutoff in degree distribution and find then the same logarithmic scaling for the diameter. However, the exponential cutoff removes the very large nodes and thus changes the character of the graph.

The graphs with $\tau > 3$ are homogeneous in the sense that, in average and asymptotically, all nodes have, so to say, the same kind of environment around themselves. This is, however, not the case one would expect for the Internet graph, where the importance of different nodes in the functioning of the network is very different — some powerful nodes have a key role. This is often reflected in their large degrees in the graph.

Therefore, it is interesting to see what happens when one leaves out the exponential cutoff, which was made only for mathematical tractability (huge degrees do indeed appear in the data, in good agreement with the power law). The effect turns out to be quite dramatic. The distance between two randomly chosen nodes shrinks to the order $\log\log(N)$, and the reason for this is that the large nodes form spontaneously a kind of core network, which provides the high connectivity. This core contains only a small (eventually vanishing) part of the nodes. Most of the nodes have a small degree, but they can be connected to the core with even a smaller number of steps than is the diameter of the core. We present the rigorous mathematical proofs only partially, the full proof will be given elsewhere. However, we treat rigorously the new crucial features which are absent for $\tau > 3$, and also make the rest of results intuitively plausible.

II. The model of Newman, Strogatz and Watts

The model is defined as follows. Let N denote the number of nodes in the graph, and let $D_1, \ldots, D_N$ be i.i.d. random variables, taking values $\{1, 2, \ldots\}$. Denote $L = L(N) = \sum_{i=1}^{N} D_i$. To make the degree sequence possible, we add one more node whose degree is

$$D_0 = 1_{\{L(N) \text{ is odd}\}}.$$

The graph is now built by joining "stubs" on the nodes randomly to form links. It may happen that some node is connected by a link to itself (a "ring"), and there can be more than one links between the same pair of nodes, but this has little significance.

Motivated by the observed power laws, we want to have

$$\mathbb{P}(D = d) \sim const \cdot d^{-\tau},$$

where $\tau \in (2,3)$. For simplicity, let us fix the distribution as

$$\mathbb{P}(D \geq d) = d^{-\tau+1}, \quad d = 1, 2, \ldots. \tag{1}$$

All our proofs assume this particular distribution.

III. Emergence of a Core When $2 < \tau < 3$

We focus on the model described in the previous section with $\tau \in (2,3)$. From a rigorous point of view, we study properties that a large graph has *asymptotically almost surely (a.a.s.)*, which means the following. Let $A = A^{(N)}$ be the event that the random graph with N nodes has a certain property. We say that A happens a.a.s., if

$$\lim_{N\to\infty} \mathbb{P}\left(A^{(N)}\right) = 1.$$

Since $\mathbb{E}\{D\}$ is finite, the usual law of large numbers implies that the total number of links is proportional to N a.a.s.:

Lemma 1: For any $\eta > 0$,

$$\sum_{i=1}^{N} D_i \in [(\mathbb{E}\{D\} - \eta)N, (\mathbb{E}\{D\} + \eta)N] \quad \text{a.a.s.}$$

Since we often base our reasoning on a given sequence of node degrees, it is good to state also the following.

Lemma 2: If $\mathbb{P}\left[A^{(N)} \,\middle|\, D_1,\ldots,D_N\right] \to 1$ in probability, then $A^{(N)}$ happens a.a.s.

We use also

Lemma 3: Let $\phi, \psi : \mathbb{N} \to \mathbb{R}$ be functions such that $\phi(N) \to \infty$, $\psi(N) \to \infty$, and that there exists a limit

$$\lim_{N\to\infty} \frac{\psi(N)}{\phi(N)} = a \in [0,\infty].$$

Then

$$\lim_{N\to\infty} \left(1 - \frac{1}{\phi(N)}\right)^{\psi(N)} = e^{-a} \in [0,1].$$

A central role is played by the following lemma which shows how large, in the sense of their aggregated number of links, two sets of nodes need to be in order to have a.a.s. a connecting link between them.

Lemma 4: Let U and V be two disjoint sets of nodes whose definition does not refer to the connection phase of the graph construction. Denote

$$F = \sum_{i\in U} D_i, \quad G = \sum_{j\in V} D_j.$$

If $FG/N \to \infty$ a.a.s., then, a.a.s, some node in U is directly connected to some node in V.

Proof. Assume, without restricting generality, that $F \le G$. It is possible to choose a sequence of numbers $\tilde{F} = \tilde{F}(N)$ such that

$$\tilde{F}(N) \le F(N), \quad \frac{\tilde{F}(N)G(N)}{N} \to \infty, \quad \frac{\tilde{F}(N)^2}{N} \to 0.$$

We show that by choosing randomly the endpoints of $\tilde{F}$ link stubs from the nodes of U, at least one of them will go to a node in V, and no one will go back to U (a.a.s.). Indeed, it is easy to check, using Lemmas 3 and 1, that

$$\begin{aligned}
&\mathbb{P}(\nexists \text{ links } U \to V)\\
\le\;& \mathbb{P}(\text{all links go to } (U\cup V)^c) + \mathbb{P}(\exists \text{ some links } U \to U)\\
=\;& \frac{L-\tilde{F}-G}{L-1}\cdot\frac{L-\tilde{F}-G-1}{L-3}\cdots\frac{L-2\tilde{F}-G+1}{L-2\tilde{F}+1}\\
&+1-\frac{L-\tilde{F}}{L-1}\cdot\frac{L-\tilde{F}-1}{L-3}\cdots\frac{L-2\tilde{F}+1}{L-2\tilde{F}+1}\\
\le\;& \left(\frac{L-\tilde{F}-G}{L-2\tilde{F}+1}\right)^{\tilde{F}} + 1 - \left(\frac{L-2\tilde{F}+1}{L-1}\right)^{\tilde{F}} \to 0.
\end{aligned}$$

■

Let us fix a function $\ell : \mathbb{N} \to \mathbb{R}$ with the following properties:

$$\ell(N) \to \infty, \quad \frac{\ell(N)}{\log\log\log N} \to 0 \quad \text{as } N \to \infty.$$

The following N-dependent definition of a "small number" turns out to be very useful:

$$\epsilon(N) = \frac{\ell(N)}{\log N}.$$

In particular, note that

$$\epsilon(N) \to 0, \quad N^{\epsilon(N)} \to \infty \quad \text{as } N \to \infty. \tag{2}$$

Now we can begin the analysis of our random graph. Let us denote the size of the largest node by

$$i^* \doteq \operatorname{argmax}_{i\in\{1,\ldots,N\}} D_i.$$

First we note that the maximum degree D_{i^*} is about $N^{\frac{1}{\tau-1}}$. Note that $1/(\tau-1) > 1/2$.

Proposition 5: The size of the largest node satisfies

$$D_{i^*} \in \left[N^{\alpha_1}, N^{\alpha_1+2\epsilon(N)}\right] \quad \text{a.a.s.},$$

where

$$\alpha_1 = \alpha_1(N) = \frac{1}{\tau-1} - \epsilon(N).$$

Proof. By (1),

$$\mathbb{P}(\max D_i < N^{\alpha_1}) \le \left(1 - \lceil N^{\alpha_1}\rceil^{-\tau+1}\right)^N \to 0$$

by Lemma 3, because, by (2),

$$\frac{N}{N^{\alpha_1(\tau-1)}} = N^{(\tau-1)\epsilon(N)} \to \infty.$$

Similarly, we see that

$$\mathbb{P}\left(\max D_i \le N^{\alpha_1+2\epsilon(N)}\right) \to 1.$$

■

It is good to note that cyclic links, connecting a node to itself, appear typically only by nodes that are bigger than $\sqrt{N}$. The proof resembles that of Lemma 4.

Proposition 6: (i) Let $\gamma \in (0, 1/2)$ and $a(N)$ be any function such that $a(N) \in (0, N^\gamma)$. An arbitrary node with degree in $[a(N), N^\gamma]$ has no cyclic links a.a.s.

(ii) Let $\gamma \in (1/2, 1)$ and $b(N)$ be any function such that $b(N) > N^\gamma$. An arbitrary node with degree in $[N^\gamma, b(N)]$ has cyclic links a.a.s.

Next we aim to show that, starting from the largest node i^*, we can reach almost all sufficiently large nodes in a relatively small number of steps. Recall the number α_1 from Proposition 5, and define recursively

$$\begin{aligned} \beta_1 &= 1-\alpha_1 = \frac{\tau-2}{\tau-1} + \epsilon(N), \\ \beta_2 &= (\tau-2)\beta_1 + \epsilon(N), \\ &\dots \\ \beta_k &= (\tau-2)\beta_{k-1} + \epsilon(N) \\ &= \frac{(\tau-2)^k}{\tau-1} + \epsilon(N)\sum_{i=0}^{k-1}(\tau-2)^i. \end{aligned}$$

Denote $U_0 = \{i^*\}$, define the sets

$$U_j = \left\{ i:\ D_i \geq N^{\beta_j + \epsilon(N)} \right\}, \quad j = 1, 2, \dots,$$

and denote

$$D_{i^*}^{(j)} = \sum_{i \in U_j} D_i.$$

Lemma 7: For any $x \in (0, 1/(\tau-2))$,

$$\sum_{i=1}^{N} D_i 1_{\{D_i \geq N^x\}} \geq \frac{1}{2} N^{1-(\tau-2)x} \quad \text{a.a.s.}$$

In particular, $D_{i^*}^{(j)} \geq \frac{1}{2} N^{1-\beta_{j+1}+(3-\tau)\epsilon(N)}$ a.a.s.

Proof. The number of nodes with positive contribution to the sum has the distribution $\mathrm{Bin}(N, \mathbb{P}(D \geq N^x))$. Since $\mathbb{P}\big(\mathrm{Bin}(n,p) \leq \frac{np}{2}\big) \leq \exp(-\frac{np}{8})$ (see, e.g., [8], (2.51)), the number of non-zero terms is a.a.s. larger than $N^{1-(\tau-1)x}/2$, and the claim follows. ■

Proposition 8: Let $j \in \{0, 1, \dots\}$. Select a node from U_{j+1} by any rule that does not refer to the connection phase of the graph construction. Such a node is connected to a node in U_j a.a.s.

Proof. Denote by D the degree of the selected node. If the node belongs to U_j, there is nothing to prove, so assume that it does not. Then it is sufficient to check the condition of Lemma 4. Indeed, we have by Lemma 7 a.a.s.

$$\frac{D \cdot D_{i^*}^{(j)}}{N} \geq \frac{1}{2} N^{(4-\tau)\epsilon(N)} \to \infty.$$

■

With respect to k, the numbers β_k have the positive limit $\epsilon(N)/(3-\tau)$. This motivates the following definition: we call the set

$$C = \left\{ i:\ D_i \geq N^{\xi(\tau)\epsilon(N)/(3-\tau)} \right\}, \tag{3}$$

where $\xi(\tau) = 7 - 2\tau$, the *core* of our random graph. (We use this term in another sense than is customary in graph theory [7], where the definition of a core refers to *subgraphs* in which the degrees of the nodes have a certain lower bound.)

Proposition 9: Let k^* denote the number

$$k^* = \left\lceil \frac{\log\log N}{-\log(\tau-2)} \right\rceil.$$

Then $C \subseteq U_{k^*}$.

Proof. In order to have $\beta_k + \epsilon(N) \leq \xi(\tau)\epsilon(N)/(3-\tau)$, it suffices that

$$\frac{(\tau-2)^k}{\tau-1} \leq \left(\frac{\xi(\tau)-1}{3-\tau} - 1 \right) \epsilon(N) = \epsilon(N).$$

Taking logarithms we see that this is equivalent to

$$k \geq \frac{\log\log N - \log \ell(N) - \log(\tau-1)}{-\log(\tau-2)}.$$

Neglecting the negative terms we obtain the expression of k^*. ■

The existence of a so called giant component (a connected component with size proportional to N) is guaranteed in classical random graphs when the number of links, chosen randomly from all pairs of nodes, is larger than $cN/2$, where $c > 1$. In our case, with $\tau \in (2,3)$, it can be shown that the giant component always exists. Although we have at present an explicit proof only for the distribution (1) with $\tau \in (2,3)$, we have good grounds to believe that the criterion for the existence of a giant component is $\sum_{i=1}^{\infty} d^2 \mathbb{P}(D=d) / \mathbb{E}\{D\} > 2$. This criterion appears in both [5] and [6], in very slightly different situations. The left hand side is the limit, when $N \to \infty$, of the expectation of the size of a node where a randomly chosen link stub is sticking out, and our argument is based on the well-known criterion for a positive probability of eternal life of a branching process. In the regime $\tau \in (2,3)$, the sum is infinite. It is clear from Lemma 4 and Proposition 5 that i^* belongs to the giant component. For $\tau > 3$, in contrast, the existence of a giant component would not depend on the distribution tail alone.

IV. Distance in a Power Law Graph with $2 < \tau < 3$

We have the following proposition, whose complete proof will be published later:

Proposition 10: Assume that the degree distribution is (1) with $\tau \in (2,3)$. Two randomly chosen nodes of the giant component are a.a.s. connected with at most $2k^*(1+o(1))$ steps.

We showed above that the core C consists of k^* layers, surrounding the node i^*, such that an arbitrary node from layer $j+1$ is a.a.s. directly connected to a node from the higher layers. One can show that the links providing those at most k^* steps up to i^* exist *simultaneously* a.a.s. — note that k^* also increases to infinity, although very slowly.

Then it remains to show that an arbitrary node from the giant component is connected to the core in at most $k^* o(1)$ steps. We skip the proof, which is based on a branching process construction. Here is a heuristic argument indicating that the core is reachable in at most k^* steps. Assume that the neighborhood reachable from a node i of the giant component in $j(N)$ steps, say $V_i(j(N))$, grows a.a.s. at least exponentially fast when $j(N) \to \infty$, $j(N) \le k^*(N)$ (note that the existing "small world" results correspond to an exponential growth of the set reachable in k hops): $|V_i(j(N))| \ge \mu^{j(N)}$ a.a.s. for some $\mu > 1$. Then a sufficient condition for $V_i(k^*(N))$ to have a.a.s. a link to the core C is, by lemmas 4 and 7,

$$\frac{|V_i(k^*(N))| N^{1-(\tau-2)\xi(\tau)\epsilon(N)/(3-\tau)}}{N}$$
$$\ge \quad \mu^{k^*(N)} e^{-(\tau-2)\xi(\tau)\ell(N)/(3-\tau)} \to \infty \quad \text{as } N \to \infty.$$

But this is true since $\ell(N)$ grows more slowly than $k^*(N)$.

V. Simulation results

The number k^* grows so slowly with N that it remains practically constant over all N that have some interest. Here are some numbers when $\tau = 2.5$:

N	10^3	10^4	10^5	10^6	10^7	10^8	10^9
k^*	3	4	4	4	5	5	5

This can be related to the empirical fact that the average hop distance in Internet has remained nearly constant while N has grown from a few thousands to about 150000 [1], [9].

We made some simulations and found that the behavior is indeed similar to what our reasoning suggested. The simulation results are summarized in Table I. They show only one simulation run for each N, so that the numbers are not statistically accurate, but give a clear impression of the overall behavior. We show the number of nodes remaining outside the giant component, the number of cyclic links, and the distances of the nodes from the largest node. The relative size of the giant component converges to a positive constant, which is in this case about 80%. The majority of the nodes is indeed connected over very few steps. In our theoretical study, we did not consider the true diameter of the giant component, but the distance between two randomly selected nodes, which is of course a smaller number. The simulations indicate, however, that the true diameter is not much bigger, and the number of exceptionally distant nodes was usually very small.

We also give two visualizations of the graphs. Fig. 1 shows all nodes and links so that the vertical axis corresponds to a node's distance from the largest node (we put distance -1 outside the giant component) and the horizontal position is the square root of the number of the node in reversed order by degree.

Fig. 2 shows all nodes as small squares ordered according to decreasing degree in the form of a quadratic spiral: 
The darkness corresponds to the node's distance from the largest node, with the same graylevels as in Fig. 1. The clearly visible outmost layer consists of nodes with degree 1. The black nodes are outside of the giant component. This figure indicates that rather few nodes with degree larger than one remain outside the giant component.

Finally, we produced a similar picture with $\tau = 3.5$, keeping the size of the giant component the same by modifying the degree distribution to be $\mathbb{P}(D \ge d) = (d+a)^{1-\tau}/(1+a)^{1-\tau}$, $a = 1.465$. The light center is clearly smaller in this regime.

TABLE I

NUMBER OF NODES OUTSIDE THE GIANT COMPONENT ("REST"), NUMBER OF CYCLIC LINKS ("RINGS"), AND THE NUMBER OF NODES AT DISTANCES 1,2,... FROM THE LARGEST NODE IN SOME SIMULATED REALIZATIONS WITH $\tau = 2.5$.

N	1001	3163	10001	31625	100000
rest	266	678	2228	6517	20913
rings	6	78	9	79	34
1	86	587	271	2691	2070
2	294	1309	2232	11644	23134
3	246	491	3421	7910	35039
4	83	88	1464	2268	14521
5	23	8	329	491	3432
6	2	1	47	84	679
7			7	14	169
8			1	4	37
9				1	5

VI. Conclusions and remarks

We have shown how a kind of "core network" arises spontaneously from the distributional assumption that the node degrees are (almost) independent and obey a power law with infinite variance. This core makes it possible that the diameter of the graph grows extremely slowly (as $\log \log N$) with its size. Such a core does not arise with $\tau > 3$ (note also that $k^* \to \infty$ as $\tau \nearrow 3$).

The practical consequence is that the existence of very large nodes has an important positive impact on the connectivity. "Small is beautiful" does not hold for architectures of large networks (although, on the other hand, this increases vulnerability for node losses — see [10]).

Properties of this kind of models have already been used in studies on denial of service attacs [11] and multicast [12], [13]; we hope that our results will turn out to be useful in that kind of "principle level" research on large telecommunication networks.

The role of the independence assumption is twofold. In the real Internet, some efforts are probably made to maintain a network that is better than a random one. Thus, the model may be motivated as a "pessimistic" one. On the other hand, the feature that the connections are made irrespective any geometric aspects can be seen as an "optimistic" bias.

There is no one-to-one correspondence between degree and power/importance of individual routers — pure core routers need not have many ports despite their high importance. On the other hand, this correspondence may be good in the domain graph, also studied in [1].

In a very recent paper [14], the same model was studied with more extensive simulations and compared with empirical data.

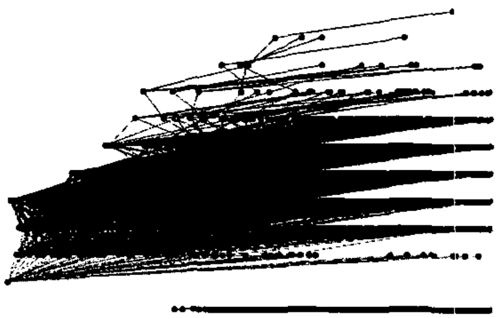

Fig. 1. Distances from largest node. $N = 5000$, $\tau = 2.5$.

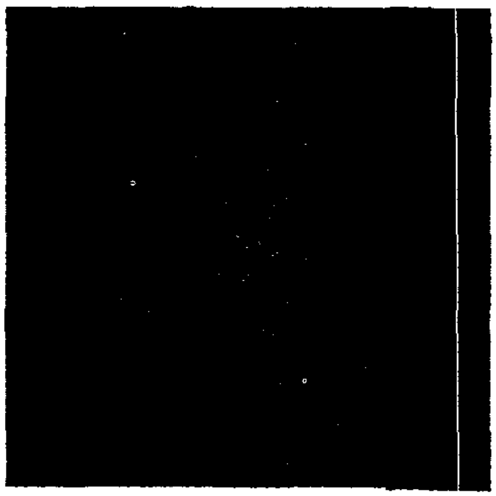

Fig. 2. 50000 nodes in spiral, ordered by descending degree. Darkness corresponds to closeness to the largest node. $\tau = 2.5$.

The model showed a surprisingly good match with the real networks. It was also possible to reveal a kind of "soft hierarchy" resembling our "core".

REFERENCES

[1] M. Faloutsos, P. Faloutsos, and Ch. Faloutsos, "On power-law relationships of the Internet topology," *ACM SIGCOMM'99 (http://www.cs.ucr.edu/~michalis/papers.html)*, pp. 251–262, 1999.

[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, pp. 509–512, 1999.

Fig. 3. Same as previous figure, but with $\tau = 3.5$, and distribution modified to keep the size of giant component unchanged.

[3] R. Albert and A.-L. Barabási, "Topology of evolving networks: Local events and universality," *cond-mat/0005085*, pp. 1–13, 4 May 2000.

[4] R.Albert and A-L. Barabási, "Statistical mechanics of complex networks," *http://arXiv.org/list/cond-mat/0106096*, pp. 1–54, 2001.

[5] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, "Random graphs with arbitary degree distribution and their applications," http://arXiv.org/list/cond-mat/0007?200, pp. 1–18, 2000.

[6] W.Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," *Proc. of ACM STOCK'2000, or http://www.isi.edu/~hongsuda/publication/index.html*, pp. 1–10, 2000.

[7] B. Bollobas, *Random Graphs*, Cambridge Univ. Press, 2001.

[8] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, John Wiley & Sons, 2000.

[9] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet map discovery," *Proc. IEEE Infocom 2000, http://www.icsi.berkeley.edu/~ramesh/papers.html*, 2000.

[10] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–381, July 2000.

[11] K. Park and H. Lee, "On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets," *SIGCOMM'01*, pp. 15–26, 2001.

[12] G. Phillips, S. Shenker, and H. Tangmunarunkit, "Scaling of multicast trees: Comments on the Chuang-Sirbu scaling law," *Proc. of ACM SIGCOMM'99, or http://www.isi.edu/~hongsuda/publication/index.html*, pp. 41–52, 1999.

[13] C.Adjih, L. Georgiadis, P. Jacquet, and W. Szpankowski, "Is the Internet fractal: The multicast power law revisited," Tech. Rep., http://www.cs.purdue.edu/homes/spa/reports.html.

[14] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network topologies, power laws, and hierarchy," *Technical Report 01-746, http://www.icsi.berkeley.edu/~ramesh/papers.html*, pp. 1–26, 2001.