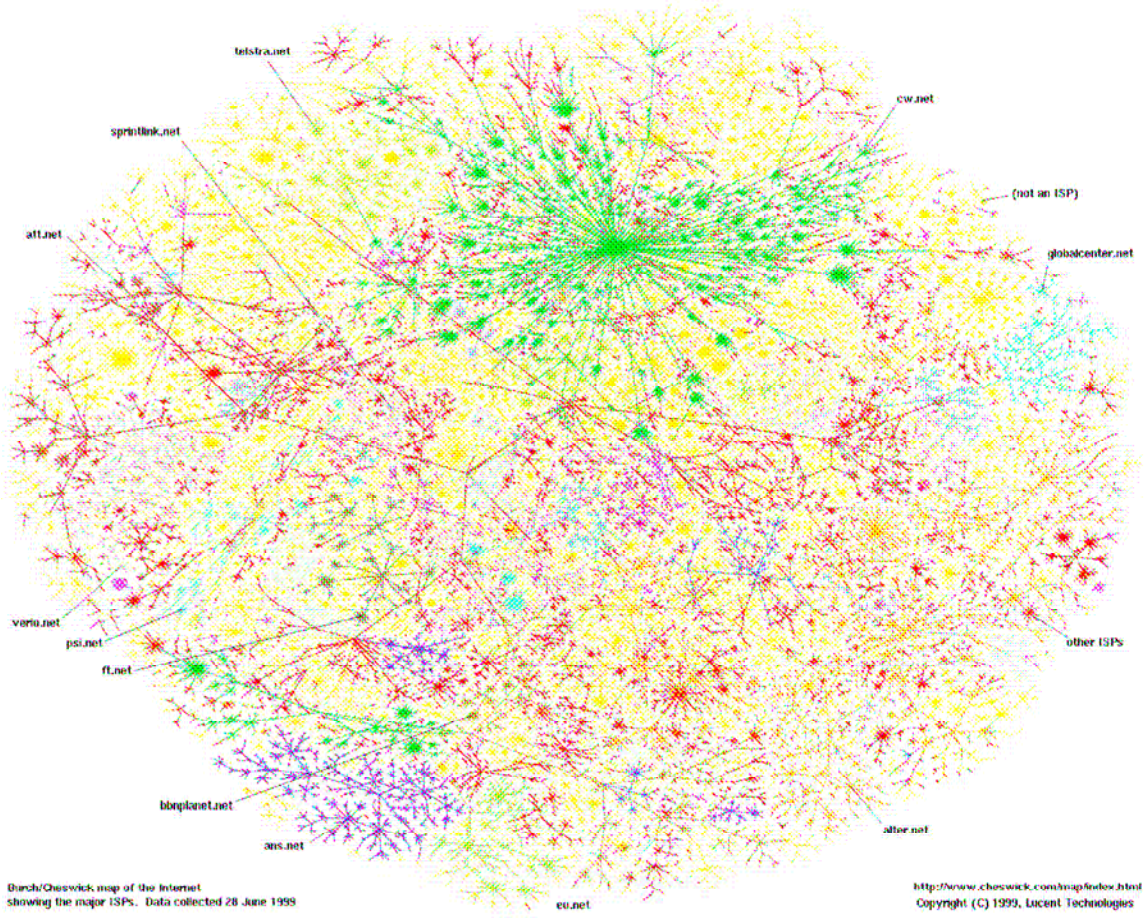
Emergent Properties of the Net

By: Guy Tal

Abstract: **The purpose of this paper is to discuss the unexpected, emergent properties of the Internet and the World Wide Web. Such properties include a fractal distribution of nodes, a scale-free power law, resiliency to breakdown under randomized attack, and threat of breakdown under specifically directed attack. Experimental evidence is surveyed and a few models are presented. The potential usefulness of such information for defense purposes and better functioning of the Internet is outlined.**[1]



[1] The picture above, from [9], is the web as seen by a single user.

Introduction

From food webs, to paper citations, to business collaborations to the Internet networks are ubiquitous. On closer inspection, these networks aren't random, nor are they all the same, a fact that has fueled network theory in the last few decades. The subject, however, is far from being merely academic, and possible applications abound.

We are continually transforming our environment in such ways as deforestation, the construction of cities, pollution, and over-fishing, and an understanding of ecological webs and food webs could help us determine how stable they are to such perturbations.

The stock market erratically rises and falls, and with it so does job security, prices, and political tensions. A detailed insight into the nature of business collaboration networks could, as in the case of food webs, help us cope with and predict the ripples caused by the collapse or construction of new companies.

The Internet is, today, central to the proper functioning of many endeavors world-wide, from scientific collaborations, to business meetings, to the dissemination of discoveries, news, and entertainment. Protecting it from intentional or unintentional attack, then, should be considered a top priority. To protect it, however, one must know its vulnerabilities, and to possess this knowledge is, in part, to grasp the properties of its network structure. Such knowledge could also be useful in optimizing the Internet's growing physical layout, or even in setting up the Internet in developing countries.

These are but three examples in a nearly endless list, and this paper will primarily focus on the last of these, the Internet and the World Wide Web. The Internet makes a good case study because, as will be elaborated on, network predictions are highly size sensitive. The bigger the network, the more applicable the theory developed becomes. And the Internet is a *big* network.

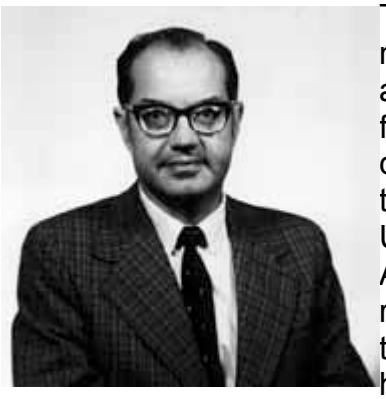
Furthermore, statistics relating to the Internet, such as the number of webpages in existence, the average number of clicks necessary to get from one webpage to another, and even the maximum of the minimum number of clicks necessary to get from one webpage to another, are much easier to gather than, say, complete knowledge of predator-prey relations in the Amazon.

While the conclusions reached in this paper are tentative, for reasons explained in the section 'Cons and Considerations,' they are tantalizing and numerous. First, the Internet is thought to be a "scale-free" network which exhibits the "small world" phenomena [1]. Second, the Internet seems robust under randomized attack but sensitive to intentional attacks [3], [4]. Also, the routers that make up its hardware are distributed in a fractal pattern which has the same dimension, within error, to the fractal population distribution [4]. Finally, and somewhat shockingly, many available topology generators used to model the Internet are fundamentally different from the current, actual Internet [2].

Though these terms are a mouthful, by the end of this paper the reader is intended to have gained a basic understanding of this jargon and the above

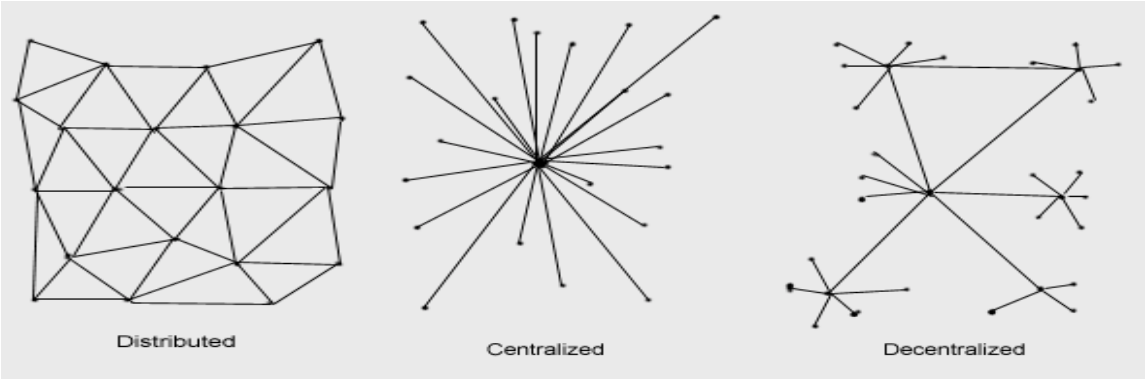
results. The derivation (where one exists) of most of these results is left to the sources.

History and Background[2]



The origins of the Internet date back to a man by name of Paul Baran (handsome fellow to the left [6]), an employee of the Rand corporation. In 1957 the former U.S.S.R. launched Sputnik which sparked U.S. concern of a possible technological gap between these unfriendly world powers. Within the year, the U.S. Department of Defense (D.o.D.) created the Advanced Research Projects Agency (ARPA) in response. The main fear shared by the members of the D.o.D. was that of communication, which was heavily phone-based. Telephones, at the time, were connected to switching offices, which were themselves connected to larger switching offices and so forth, with little redundancy. What would happen to communications, military officials asked, if a few central offices were attacked in a nuclear war?

In 1960, the U.S. Air Force contracted Baran to work on the dilemma. Baran realized that the crux of the problem was that the two existing sorts of networks were the ones pictured below [6], called centralized or decentralized networks. Centralized networks are networks whose branches all connect to a single hub, while the decentralized networks of the time were nothing more than a few centralized networks linked together.



Baran's first insight was that a distributed network (illustrated above [6]) would be much more robust to attack. This intuition is tested in [2] and [3] and the results are presented in the section 'Experiments and Results.'

Baran's second insight related to the method of transmission. Since his network was originally to be implemented with telephones, he faced a new problem; analog signals could not be transmitted over the long distances where

[2] All history is taken from [6] and [7], and Wikipedia .

the switching offices were placed.[3] He devised the so-called digital "packet" switching system, whereby a signal was digitized and sliced into packets. Each packet was then individually sent to whatever node the system determined was the quickest (at that instant) means to the final destination. The receiving nodes, in turn, would calculate anew which available node was now the quickest. When all packets were received, the signal would be reassembled. [4]

A few years after Baran published his results, Larry Roberts, then director of ARPA, decided to implement Baran's scheme to facilitate communications among ARPA researchers, and ARPANET, the granddaddy of the Internet, was born. 20 years later, through various transformations, the Internet was fully formed. But it was only after the National Center for Supercomputing Applications, *here at UIUC*, released the Mosaic web browser version 1.0 in 1993 that the Internet went from being an academic tool to the popular entertainment pastime it is today!

Terminology[5]

To understand the results relating to the Internet one must first become familiarized with the associated terminology.

First there is the central notion of a ***graph*** or ***network***, used interchangeably. A graph consists of ***nodes*** or ***vertices*****,** denoted by dots in the above pictures of networks, and ***edges*** or ***links*****,** denoted by lines connecting nodes. Each edge is usually assumed to have length 1.

A ***directed graph*** is one in which each edge is an arrow allowing movement in one direction. The World Wide Web is a directed graph, in that one website can link to another, but the second website may not contain a link back. The Internet, however, is an undirected graph, insofar as traffic can flow either way.

An important property of a node is that of ***degree***[6]. The degree, k, of a node is the total number of links it has. In a directed graph, ***out-degree*** and ***in-degree*** are often specified, being nothing more than the total number of edges directed out of or into a node, respectively. The degree of the node is then the sum of these. A ***degree distribution***, P(k), refers to the distribution that, statistically, determines the degree of all nodes of a network.

If the degree distribution of a network follows a power law of the form k^{-v}, where v is a constant, then the network is said to be ***scale-free***. On a log-log plot of P(k) vs. k one would see a linear relationship between the two variables on "all scales" or for all values of k, leading to the name of the term.

---

[3] The distances between switching offices were far because they were specifically not placed near each other nor near population centers, both scenarios thought to create obvious targets.

[4] It's worth noting that Baran felt his network had the added virtue of being a nuclear deterrent. He argued that if a country did not feel its communications were so vulnerable to sudden attack it would be less likely to implement a first strike.

[5] The terminology presented closely resembles that of [1].

[6] Sometimes the word ***connectivity*** is used instead, though this word has several meanings in the field which should not be confused.

Another useful vertex property is its ***clustering coefficient***. This is the number of present edges between a particular vertex's nearest neighbors, divided by the number of possible edges between these neighbors. Averaging over all vertices of a network one gets the clustering coefficient of the network. "The cluster coefficient of the network reflects…the extent to which the nearest neighbors of a vertex are the nearest neighbors of each other" [1]. In other words, "the "cliqueishness" of the mean closest neighborhood of a network vertex" [1].[7]

Next, the ***distance,*** L_{xy}, between two vertices x and y, is defined to be the shortest-path length between them. The average distance, <L>, is known as the ***diameter*** of the network[8]. The diameter effectively determines the "size" of the network.

A network with a significantly larger clustering coefficient and a significantly smaller diameter, as compared to a random graph with the same number of vertices and edges, is called a ***small world***[9] network.

Alternatively, one can consider the ***maximal shortest-path length*** over all pairs of vertices between which a path exists. This is a measure of "the maximal extent of a network" [1].

Moreover, there are the notions of ***equilibrium*** vs. ***non-equilibrium*** networks, best illustrated by examples from [1].

"An example of an equilibrium network: a classical[10] undirected random graph defined…by the following rules:

i) The total number of vertices is fixed.
ii) Randomly chosen pairs of vertices are connected via undirected edges.

…The example of a non-equilibrium random network: A simple random graph growing through the simultaneous additions of vertices and edges. Definition of this graph:

i) At each time step, a new vertex is added to the graph.
ii) Simultaneously, a pair (or several pairs) of randomly chosen vertices is connected."

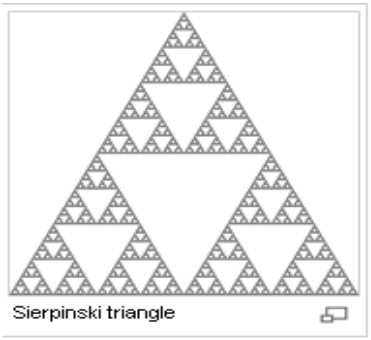
It should be intuitively clear that the first example tends to "equilibrium" configurations because the number of vertices is fixed. On the other hand, the second configuration favors older vertices which will, statistically and persistently,

---

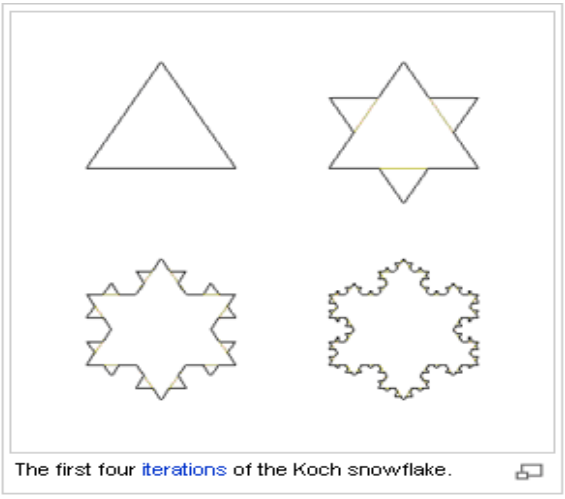
[7] This term was originally coined in a sociological context.
[8] The average is taken with respect to all pairs of vertices with an existing path length between them and all realizations of a network with a given P(k).
[9] Think: "Six Degrees to Kevin Bacon." This is a phrase originating in the famous psychology experiments of Stanley Milgram which attempted to demonstrate that communities in the world are highly connected.
[10] Vertices are statistically independent and equivalent.

have more connections than new ones, thus being in "non-equilibrium."[11] In reality, most interesting networks, such as any mentioned in this paper, are of the non-equilibrium variety, though many derived results, due to mathematical difficulties, are proven for equilibrium conditions.

Finally, there are the entwined notions of a ***fractal*** and its ***box-counting dimension***, Kenneth Falconer once said that, "the definition of a 'fractal' should be regarded in the same way as the biologist regards the definition of 'life'. There is no hard and fast definition…just a list of properties characteristic of a living thing…though there are living objects that are exceptions to each of them." The generic properties usually ascribed to fractals are those of self-similarity[12] (perhaps approximate or statistical), detail on arbitrarily small scales, and a non-integral box counting dimension. Examples of fractals are the Koch snowflake, and the Sierpinski triangle below (from Wikipedia).[13] It is important to realize that physical fractals are always "fractal" within a bounded range of sizes.



Sierpinski triangle



The first four iterations of the Koch snowflake.

The ***box counting dimension*** of a set, like ***equilibrium***, is also best illustrated by example. Consider a square, A, of side length 1. If I were covering it with other squares of side length 1, I would need only one of these covering squares to cover square A. If instead each covering square has side length ½ I would need at least 4 of them to cover square A. If each covering square has side length ¼ I would need 16 of them to cover square A, and so forth. In general, if a covering square has side length $X = (½)^n$ the number of covering

[11] As the reader may have guessed a statistical mechanics of networks is being developed.

[12] Notice how the Sierpinski triangle is made of 3 identical, shrunken versions of itself. This is an example of "self-similarity."

[13] Fractals have all sorts of unusual properties. For example, note that the ratio of each iteration of the Koch snowflake to each preceding iteration is 4/3. This implies that as the iterations tend to infinity the length of the snowflake becomes infinite even though its area is clearly finite! For this reason, Mandelbrot has observed that the lengths of rugged coastlines approach infinity as the ruler used to measure them is made smaller and smaller.

squares necessary to cover square A is $N = 4^n$. Given these relationships, it is easy to see that $\log_2 N = -2\log_2 X$, or simply that $N = X^{-2}$. As it turns out, the exponent is exactly the usual dimension we would grant a square! Experimenting with this procedure on various other simple objects results in the exponent always giving the desired dimension. It then seems reasonable to define the box counting dimension of any set, A, as the exponent, D, in the proportionality between N and X^{-D}—where N is the minimal number of boxes in the embedding space of A needed to cover A, and X is the boxes' side length—if such a proportionality exists. This aspect of dimension captures fairly accurately how much of the embedding space a particular set occupies. For example, it can be shown that the dimension of the Koch snowflake is approximately 1.262, corresponding to the intuition that it's more "space-filling" than a line segment because of its infinite extent and finite area, but less space-filling than a square.

With these concepts at hand, the reader now has the ability to not only comprehend the conclusions about the Internet that are to follow, but also to benefit from a perusal of many articles in the field of network theory.

Experiments and Results

What is the Internet and is it synonymous with the World Wide Web? Though the two terms are often used interchangeably a distinction exists. The Internet is, roughly speaking, the "hosts (computers of users), servers (computers or programs providing a network service that also may be hosts), and routers that arrange traffic…The routers are united in domains." [1] In other words, the Internet is the "interconnected computer networks linked by copper wires, fiber optic cables, wireless connections, etc."[14] On the other hand, the World Wide Web is "the array of its [the Internet's] documents plus hyperlinks" [1]. As mentioned earlier, the Internet is an undirected graph while the World Wide Web is a directed graph so the distinction is not merely pedantic.

Thus there are at least three networks one can consider: the Internet on an inter-domain level, the Internet on a router level, and the World Wide Web, and this paper will now expound on what is known of the details of the topologies of these various networks and how successful modeling them has been.

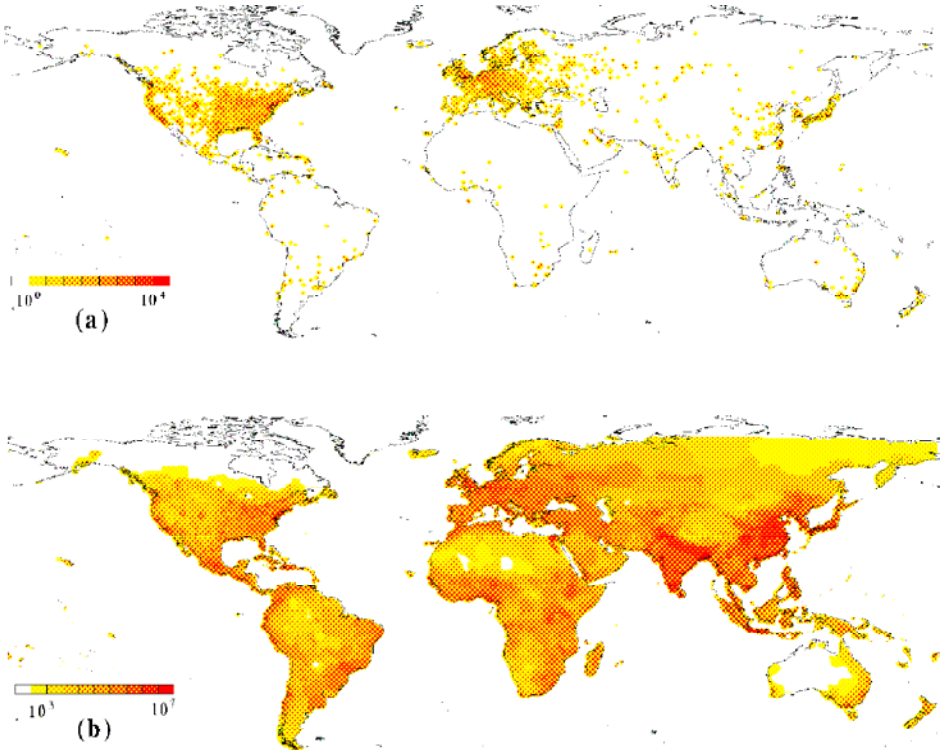
On the router or domain levels, the distribution of the Internet around the world is found to be fractal in nature.[15] [2] As one might suppose, the higher the population density of a region of a technologically developed country, the more Internet demand there is from the population. This deduction is visually verified in the figure below [2] where both the fractal nature of the router density (a) and the fractal nature of the population density (b) of North America are presented.[16] The dimension of all three fractals (including domain-level) is empirically, and strikingly, found to be 1.5 with an error of .1.

---

[14] Quote from Wikipedia with "Internet" as a search term.

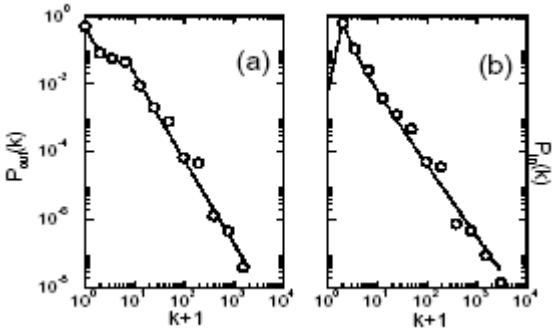
[15] The bounds of the fractal approximation are naturally constrained between the single router and the entire network.

[16] Both maps use a box resolution of $1^o x 1^o$. The gradients on the lower left indicate people/box. As for the scale of the statistics, note that 228,265 router coordinates were used.



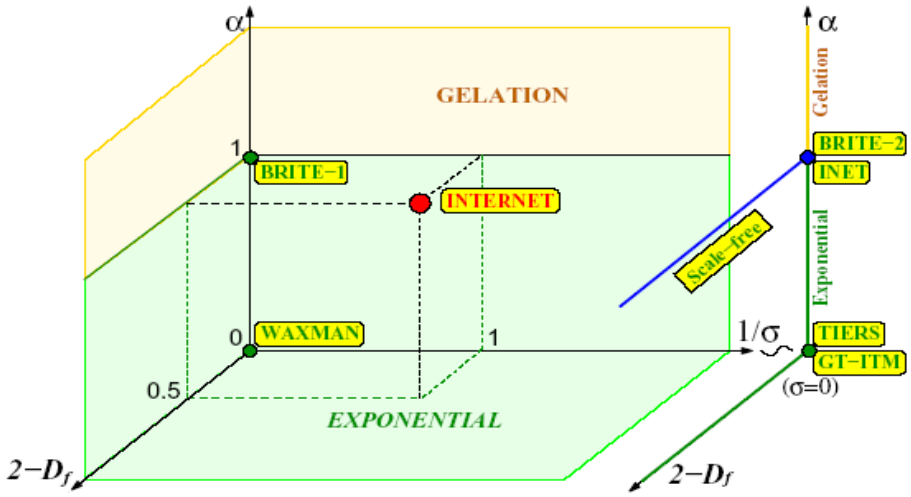
This fractal distribution of the nodes of a network is in contrast with the traditional network paradigm wherein nodes are randomly distributed, possibly giving rise to errors in current Internet topology simulations.

Moreover, [1] claims that on all levels the Internet is “scale-free,” with an exponent between 2 and 3. [5] determines the exponent of the WWW to be 2.1 and 2.45 for incoming and outgoing edges respectively, from the slopes of the lines below.



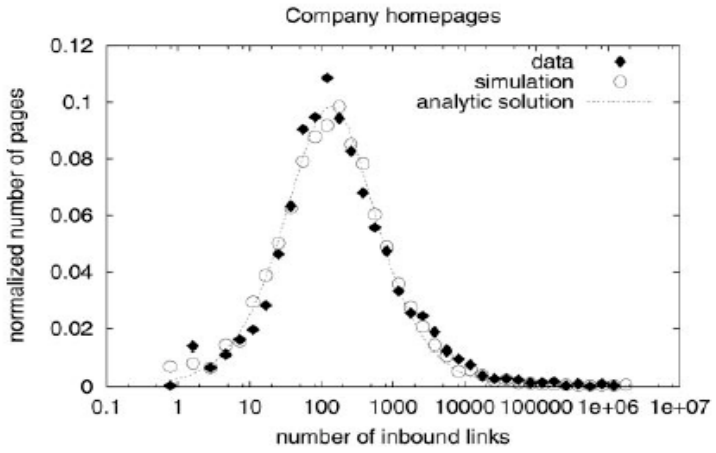
The usual explanation for “scale-free” behavior is preferential attachment, which means that a new node is more likely to attach to a node with more links than to one with less, otherwise known as “the rich get richer” scheme. [2] elaborates on this description and concludes from measurements that link placement is actually determined by *two* competing mechanisms. “First, the likelihood of connecting two nodes decreases linearly with the distance between

them, and second, the likelihood of connecting to a node with k links increases linearly with k." Using the parameters (ά, σ, D_f) where ά and σ are the exponents that govern preferential attachment and the cost of node-node distance, and D_f is the dimension, [2] plots out, from measured values, the precise location of the Internet's large scale topology in this phase space, along with the location of several other known topology generators. The results are shown below.



As can be seen, the Internet occupies a distinctly different position of the diagram than any generators shown purporting to model it, thereby debunking their results.

However, [8] points out that even this is not the full story. If we were to discard the main hubs of the web, and focus our attention instead on smaller communities of interest, such as all university or newspaper homepages, the scale-free nature would disappear, replaced by a unimodal distribution, illustrated below for company homepages.



On a related, and final, note, since scale-free networks have hubs with high degree, due to preferential selection, the diameter of such networks is

typically small. Combined with the fact that these networks tend to have large cluster coefficients this implies that scale-free networks usually exhibit the small world phenomena. This is verified for the World Wide Web in [5], which measures the diameter of the web to be <L> = .35 +2.06log(N), where N is the number of documents available on the web. At the time of [5]'s writing[17], the web had approximately 8 x 10^8 documents, giving a diameter of 19 links. Even if we scale this up by 1000% we find the diameter to only have increased to 21 links, making the World Wide Web a small world indeed.[18]

To summarize, it appears that the Net[19]:

- Develops in fractal fashion, driven by the fractal population distribution.
- Exhibits small world, scale-free behaviors due to competing drives between preferential selection and node-node distance constraints.
- Has a unimodal distribution, rather than a power law, when considering certain subnets.

The next step for researchers, as always, would be to incorporate these newer discoveries into even more complex Net topology generators and see how our growing understanding fares with our observation. Since the Internet is only about 20 years old, and this field of study is younger still, there is ample reason to suspect that we have not yet fully captured the depth of this captivating network.

Cons and Considerations

Curiously, the above 3 tendencies, for the web imply that it is both robust and fragile. [4] has analytically shown that scale-free networks with exponents between 2 and 3 are impressively stable under random collapse of nodes, never disintegrating unless the network is finite, and approaching perfect stability the larger the network.[20] The Internet, then, is effectively indestructible in this way. It is not difficult to see why this is the case. The small world aspect of the Internet guarantees that there be multiple ways from point A to point B on a network. This is nothing more than Paul Baran's original insight into distributed networks.

However, [3] has shown that on scale-free networks intentional attack—where a fraction of the most connected sites are suddenly removed—can be calamitous, and that even well before the threshold of total collapse, there are noticeable effects.

In a sense, this is the same problem that the D.o.D. originally faced with the decentralized phone networks, suggesting that the practical realization of

---

[17] Sept. 10, 1999.

[18] This should give hope to any thoughtful web searchers. The site you're looking for is only a few clicks away!

[19] I use this term to collectively refer to the 3 networks discussed.

[20] Technically, it is not the entire network that is stable, it is the spanning cluster, because initially there could have been a few small disconnected parts. Loosely speaking, the spanning cluster is the largest connected subnetwork of the original network. For a more rigorous approach see [1].

Baran's solution to the nuclear threat on communication, or what's become of it, is not entirely as sound as it seems.

On top of that, the web's small world tendency makes epidemic computer viruses much more common. On the flipside, it also makes the dissemination of knowledge equally epidemic. As with most powerful inventions, the Net's complex structure can be used for both good and ill.

It is important, however, to remember the assumptions implicit in the above results. First, [1] continually emphasizes the noticeable finite size effects on experimental data sets, and the difficulty they create in extracting a definitive power law. Also, collecting such data as the maximal shortest-path length is non-trivial and may be done incorrectly. Second, there is the subtle point that since the Net isn't in equilibrium, and since its uses change with time, its topology might as well. Since we haven't studied the Net for very long, and since most studies reference data collected a few years earlier, this may have already begun without our noticing.

Two effects that might play a role in such a change are the aging of nodes and the cost of adding links to or limited capacity of a node. The aging of nodes refers to the situation where, because of their age, nodes begin to lose popularity rather than continue to gain it, thereby altering the previously scale-free nature of the network. The potentially limited capacity of a physical node may also contribute to truncation of the scale-free phenomena. Both effects, if present, have been shown to alter a network's power law distribution [9].

The former effect has been observed in networks of actors, where the aging of an actor literally corresponds to the actor's getting old and, as a result, ceasing to gain popularity. [9] In the web, this could correspond to a fad running its course, or a marketing campaign losing its appeal to a new generation.

Finally, [3] assumes that the Internet is in equilibrium, but this isn't true. At best, this approximates the Internet at an instant in time. Not only is the Internet growing, but it is constantly interacting with its other half, the World Wide Web. The analysis needn't stop there. The users of the World Wide Web are a part of the food web. The food web is in constant interaction with the network of political alliances, and so and so on. Depending on the scope of analysis, there are potentially endless numbers of networks interacting with networks, like microorganisms in a primordial soup, evolving into an ever more intricate, emergent web.

Citations

1) Dorogovtsev, S. N., Mendes, J.F.F. Evolution of networks. Advances in Physics (2001).
2) Yook, S. -H., Jeong, H. and Barabasi, A.-L., 2001, Modeling the Internet's large-scale topology.
3) Cohen, R., Erez, K., Ben-Avraham, D. and Havlin, S., 2000, Resilience of the Internet to random breakdown. Phys. Rev. Lett., 85, 4625 (2000)

4) Cohen, R., Erez, K., Ben-Avraham, D. and Havlin, S., 2001, Breakdown of the Internet under intentional attack. Phys. Rev. Lett., 86, 3682
5) Albert, R., Jeong, H. and Barabasi, A. –L., 1999, The diameter of the world-wide web. Nature, 401, 130.
6) Griffin, Scott. “Internet Pioneers.” http://www.ibiblio.org/pioneers/baran.html.
7) Vuyk, Brian. “The Influence of Paul Baran on the Development of the Internet,” 2006, http://www/infohatter.com/node/4.
8) Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L., Winners don’t take all: Characterizing the competition for links on the web. Proceedings of the National Academy of Sciences, 99(8);5207-5211, 2002.
9) Caldarelli, G., Marchetti, R. and Pietronero, L., 2000, The fractal properties of Internet, Europhys. Lett., 52 386.