TRN's
Making the Future report
The State of an Emerging Technology and a Look at What Lies Ahead
Report Number 14 May, 2004

Internet Structure: The Global Nervous System

Executive Summary

Measuring and mapping the Internet is challenging because the global network has many dimensions and is so very large — more than 230 million servers and billions of Web pages that contain hundreds of terabytes of data.

This size also holds the key to the network's deeper nature. Because it is too big for any one group or organization to design or control, its structure is governed by natural laws. It is dynamic, with change its critical dimension and growth its driving force.

The Internet is really a pair of superimposed networks — one physical, the other virtual.

Researchers looking for growth patterns in the network of servers, routers and network connections that make up the physical Internet are finding that the Internet has slowly shifted from a highly distributed structure to a somewhat more centralized structure, and because of this it is more vulnerable to attacks from hackers, terrorists and errant backhoes.

Researchers are looking to exploit the structure of the virtual Internet — the World Wide Web — to improve searching and browsing, and to shore up the network against attacks from viruses and worms.

The Internet also serves as an easily-measurable and observable network that can inform scientists about many other networks, including social networks, chemicals used in life processes, and mobile phone connections.

Researchers have discovered that the Internet is scale-free, with a few large sites and many smaller sites, and small-world, with many shortcut links among interlocking groups.

This makes it relatively easy to get from one point to any other on the Net, and it's likely to stay that way even as the Web grows larger. The same attributes, however, make the network more vulnerable to attack from viruses and worms, and fairly resistant to inoculation.

What to Look For

Search and Browsing:

Crawling and ranking methods that correlate text and links
Custom search engines that leverage link structure
Peer-to-peer searching methods that tap link structure
Searching methods that leverage small-world structure
Content filtering that taps link structure
Discoveries about how people use link structure cues

Optimization:

Traffic flow optimization that leverages link structure
Detailed Internet simulations for optimization
Distributed caching schemes that leverage user clusters

Fault Tolerance:

Network design methods that tap link structure
Improved ways of estimating the impacts of node failure
Internet immunization strategies that leverage link structure
Detailed Internet simulations that better identify weak spots

Analysis and Theory:

Measures of link competition within communities
Improved analysis of Web clustering
Scientific classification of networks

The elephant and the blind men

The Internet is a massive, sprawling, dynamic network of computer networks. It is also a manifestation of social, political and economic relationships. Measuring and mapping the Internet is challenging because it has many dimensions and is so very large — more than 230 million servers and billions of Web pages that contain hundreds of terabytes of data.

Although it is relatively easy to estimate the Internet's size, getting a sense of the Internet's structure is more difficult. Existing maps show how the Internet's subnetworks are connected, and there have been various attempts at mapping the links among Web sites. But capturing the dynamics that govern the Internet's growth and determine its structure is an emerging science.

This report covers the latest research them aimed at developing methods of measuring and understanding the structure of the Internet. It also discusses the results of these measurements, which show that, at the highest level, the Internet is neither random nor designed, but is evolving according to many of the same rules that govern large natural networks.

Discovering the details of Internet structure is important because its structure governs many important aspects of the Net:

- How it continues to grow
- How easy it is to search
- How easy it is to browse
- How easy it is to organize
- How vulnerable it is to physical attacks
- How vulnerable it is to software-based attacks
- How responsive it is to attempts to inoculate it against attacks
- How well it recovers from attacks

Networks all around

A network is simply a group of elements, or nodes, that are linked in some way, and the world is full of them: the brain's billions of connected neurons, the millions of connected computer components on chips and circuit boards, social networks of people who know or work with each other, relationships among a cell's biological processes, and even the relatively fleeting connections among mobile cell phones.

The Internet is really a pair of superimposed networks — one physical, the other virtual. The physical network is made up of hardware: the servers that store the information contained within the Net are nodes and the physical connections that allow for information transport are links. The virtual network is the World Wide Web — a dynamic network of Web pages connected by links among pages.

Internet structure has several dimensions in the physical realm and several dimensions in the virtual realm.

The physical Internet:

- Servers — the computers that house content
- Routers — the computers that link the physical Internet and govern its traffic

The virtual Internet — the World Wide Web:

- Content — the text, graphics, audio and video contained in Web pages
- Links — the connections between Web pages
- Hops — a measure of the link distance between a pair of Web pages

One major difference among types of networks is how often links change. Circuit board links are largely permanent,

How It Works

Search engines find content via spiders that go through the pages on a Web site by following the links among pages. This information is stored in an index that is used to match query terms.

There are two challenges to making Web searches available this way. The first is dealing with the sheer size of the Internet. The second is being able to present a user with a reasonably whittled-down number of useful links.

Scope

The Internet is big by any measure:

The Internet Systems Consortium pegged the number of Internet hosts as of January, 2004 at 233 million.

Global Reach counted 729 million users online as of March, 2004.

And a University of California at Berkeley study showed that, in 2002, 532,897 terabytes of new data flowed across the Internet, 440,606 terabytes of email was sent, and the Web contained 167 terabytes of data that was accessible to all users, plus another 91,850 terabytes in the deep Web where access is controlled.

A terabyte is 1,000 gigabytes, or 1,000,000 megabytes, or the amount of information that can be stored on 213 DVDs, or one tenth the amount of information stored in the entire Library of Congress print collection.

Limits

This is a lot of information, and it lives in a world in which computers are only so fast and hold only so much information. There is simply not enough time and compute power for spiders to crawl all the information in anything like a timely manner, or for even the tens of thousands of servers deployed by the major search engine companies to index and cache it.

To get around the problem, today's search engines cover only 10 to 20 percent of the Web, and even then, spiders take weeks to finish a single crawl of just that portion. Search engines often crawl popular sites more often to keep them more up to date, but in general, when you search the Web or access a search engine's cached copy of a page, you are working with a snapshot that is days or weeks old.

Link structure

Link structure already plays an important role in the second challenge for search engines — presenting links that are relevant. And it is starting to play a more important role in the first challenge — covering more of the Web.

biological processes change very slowly through mutations, the physical connections among Internet servers change more quickly, and links among Web pages change even more quickly. Even Web page links are relatively stable when compared with mobile phone connections. These differences affect the way these networks grow.

Because the Internet is a technological development, and because it is constrained by the corporations that own much of its infrastructure and the governments that regulate its use, it is difficult to think of the global communications network as an organic, evolving system subject to laws of nature. But because its growth is largely beyond direct human control, the Internet exhibits many of the characteristics of the big networks found in nature.

Its accessibility and digital nature makes the Internet a ready-made laboratory for researchers to not only discover the Internet's secrets, but also gain more knowledge about networks in general and compare different types of networks.

Researchers are studying the Internet directly and also by determining how close simulations come to mimicking real Internet structure.

Geography, politics, economics and fractals

Researchers are looking for patterns in the growth of the network of servers, routers and network connections that make up the physical Internet in order to follow its progress, learn about networks, and find ways to improve Internet efficiency.

In general, they are finding that the Internet has slowly shifted from its original, highly distributed structure to a somewhat more centralized structure, and because of this it is more vulnerable to attacks.

Researchers from the Abdus Salam International Center for Theoretical Physics in Italy, the International School for Advanced Studies in Italy and the Polytechnic University of Catalonia have analyzed three years worth of daily maps of the Internet's connections, and found that despite its high-growth rate, the Internet has settled into a state whose overall topographical and geometrical properties are stationary. (See "Internet Map Improves Models", page 9.)

Researchers from the Nordic Institute for Theoretical Physics in Denmark, Brookhaven National Laboratory, the Niels Bohr Institute in Norway, and the Norwegian University of Science and Technology have found that the Internet has an underlying modular structure regulated by the number of nodes, or sites, that link to a given node. The Internet has about 100 modules that correspond roughly to countries, and the farthest points from each other are Russia and U.S. military sites. (See "Study Reveals Nets Parts", page 10.)

A University of Cincinnati researcher has found that the commercial Internet has shifted from its original distributed structure toward a hub-and-spoke topology similar to those airlines use to plot routes. The shift has made the network more vulnerable to attack in the same way a major hub city is

Perhaps the best known example of using link structure to determine link relevance is Google's PageRank algorithm, which orders search results using an algorithm that measures a page's popularity based on the number and status of pages that link to it.

PageRank assigns a value to a page by adding up the values of its inbound links. A link's value is determined by the originating page's value divided by the number of its outbound links. The algorithm aims to identify authoritative sources and use their authority to evaluate other sources. Because pages determine each other's rankings, the algorithm has to run many times before it converges on a reasonable value for a given page.

Clusters

More recently, researchers have been using link structure to categorize the Internet by subject in order to identify portions of the Web that are more manageable than the entire thing. Given that pages are likely to link to related pages, search algorithms can be tuned to find densely interconnected communities of interest.

Who to Watch

Links and hops

Lada Adamic, Hewlett-Packard Laboratories
Palo Alto, California
www.hpl.hp.com/personal/Lada_Adamic

Albert-László Barabási, University of Notre Dame
Notre Dame, Indiana
www.nd.edu/~alb

Shlomo Havlin, Bar-Ilan University
Ramat Gan, Israel
ory.ph.biu.ac.il/~havlin

Jon Kleinberg, Cornell University
Ithaca, New York
www.cs.cornell.edu/home/kleinber

Alessandro Vespignani, Abdus Salam International Centre for Theoretical Physics
Trieste, Italy
www.ictp.trieste.it/~alexv

Duncan J. Watts, Columbia University
New York, New York
www.sociology.columbia.edu/people/professors/djw24/index.html

Content

Gary William Flake, Yahoo! Research Labs
Pasadena, California
labs.yahoo.com/~flakeg

C. Lee Giles, Pennsylvania State University
University Park
clgiles.ist.psu.edu

vulnerable to bad weather, which can affect flights all over the country. (See "Hubs Increase Net Risk", page 11.)

University of Notre Dame researchers have found that Internet routers are physically arranged in a fractal pattern. Fractals repeat at many size-scales and can be represented by mathematical formulas. The researchers showed that the pattern of Internet router locations that shows up on a map of the entire Internet is also discernible on maps of various-size portions of the Internet. (See "Net Devices Arranged Fractally", page 13.)

Fractals are present in many systems, including fluctuations in the stock market, the distribution of galaxies in the universe, the turbulent flow of fluids, biological growth, and coastlines. The router fractal pattern is distinct from the virtual link structure of the Web, which also follows a fractal pattern. Fractal systems are nonlinear, which makes them difficult to predict and very sensitive to small changes.

Filippo Menczer, University of Iowa
Iowa City, Iowa
dollar.biz.uiowa.edu/~fil

Comparisons

Uri Alon, Weizmann Institute of Science
Rehovot, Israel
www.weizmann.ac.il/mcb/UriAlon

Hawoong Jeong, Korea Advanced Institute of Science and Technology
Daejeon, Korea
physics.kaist.ac.kr/english/hawoong_jeong.htm

Mark Newman, Santa Fe Institute
Santa Fe, New Mexico
www.santafe.edu/~mark

Links

The structure of the virtual Internet — the Web — is determined by the ability of any page to link to any other page. Several rules have emerged that explain how the Internet is structured and grows in terms of the links between Web pages:

- A scale-free, or power-law, structure
- Small world, or six-degrees-of-separation link connections
- Short paths among links
- Rich-get-richer link growth

In scale-free networks, a few nodes, or servers, have many links to other nodes, while many nodes have relatively few links. The Internet shares this characteristic with many types of networks. (See "Network Similarities Run Deep", page 38.)

In small-world networks, a user can get from one node to any other node by traversing only a few links among adjoining nodes. This is a network trait shared by social networks.

The small-world phenomenon was first discovered by sociologist Stanley Milgram in a 1967 postal experiment. The six-degrees-of-separation cliché was spawned after Milgram found that it took an average of only six exchanges, or hops among acquaintances for a letter to find its way from a random correspondent in Nebraska to a Massachusetts recipient identified only by a brief, vague description. But only recently have researchers studying the Internet been able to explain why the phenomenon occurs.

Short paths are shortcuts that make it easy to get from anywhere in the network to anywhere else. A network has a small-world structure if a significant proportion of its links connect distant parts to make these short paths. A network's degrees-of-separation has to do with the number of short paths it contains, which can be measured as the average minimum number of links between nodes. This is also referred to as its diameter.

The Web also follows a rich-get-richer dynamic that researchers have shown contributes to its scale-free nature: the larger a node is, the more likely it is to attract links.

Link patterns

Researchers looking to map the Web's link structure more closely are discovering new phenomena.

Researchers from the University of London have discovered a structural element of the Web dubbed the rich-club phenomenon: large, well-connected nodes have more links to each other than to smaller nodes, and smaller nodes have more links to the larger nodes than to each other. The researchers found that 27 percent of links are among the largest five percent of nodes, 60 percent connect the remaining 95 percent to the largest five percent, and only 13 percent are between nodes that are not in the top five percent. The study suggests that the Internet is more dependent on the larger nodes and thus more vulnerable to attacks than previously thought. (See "Big Sites Hoard Links", page 14.)

NEC Research Institute researchers have found that the structure of specific Web communities can be different from the structure of the Web as a whole. Across the Web, nodes that already have many inbound links tend to receive even more

over time — the rich-get-richer phenomenon. The researchers have showed that within specific Web communities this can vary. The community of e-commerce Web sites selling publications, for instance, is dominated by Amazon, is highly competitive, and is structured similarly to the Web as a whole. The community of e-commerce Web sites selling professional photographer services, however, allows smaller sites a better chance of gaining links in order to grow. (See "Odds Not Hopeless for New Web Sites", page 14.)

Researchers from Kongju National University in Korea and the Korean Electronics and Telecommunications Research Institute have devised a mathematical model of the Web that shows that the rate of increase in the number of Web sites influences the pattern of Web growth. One effect of this phenomenon is that the faster a segment of the Web grows, the sharper the distinction between its large and small sites — the haves and have-nots of the Internet. (See "Faster Growth Heightens Web Class Divide", page 15.)

Researchers from the Niels Bohr Institute in Denmark have found that Web simulations more accurately follow the structure of the real-world Web when they take into account the potential for waning popularity as a Web page gets old. The age variable plays a part in the stability of real-world networks, according to the researchers' work. (See "Page Age Shapes Web", page 16.)

Link dynamics

Researchers are also trying to tease out the relationship between the way links change and the structure of the Web.

A researcher from St. Petersburg State University in Russia has found that the Internet's scale-free nature could have something to do with link dynamics. Her disappearing link model shows that when the rate of link appearance and disappearance is around 0.5 percent it causes a scale-free structure, and it shows that the rate is probably more important in determining network dynamics than network size. The 0.5 percent rate is close to the link appearance and disappearance rate in both the Internet and in the social network of scientific paper citations. (See "Disappearing Links Shape Networks", page 17.)

A researcher from the Jozef Stefan Institute in Slovenia has put together a five-million node simulation of the Web that analyzes how frequently links are updated and how outgoing and incoming links are related. The model suggests that the shape of the Web is largely determined by how often and how Web pages are updated. (See "Simulation Sizes up Web Structure", page 18.)

Getting there in fewer hops

The Internet's small-world nature assures that it's relatively easy to get from one point to any other, and it's likely to stay that way.

Researchers from Bar-Ilan University in Israel have found that the average number of connections needed to get from one point to another in real-world networks like the Internet and social networks is smaller than the number needed for randomly-connected networks. (See "Net has Few Degrees of Separation", page 19.)

Researchers from Stanford University have constructed a model of a scale-free, small-world network with short paths and used it to show that as networks like the Internet that harbor these three traits grow very large, the number of hops a data packet takes to get from one node to another node shrinks. (See "Internet Stays Small World", page 20.)

Researchers from Columbia University and the Santa Fe Institute have constructed a mathematical model that shows that groups are key to the six-degrees-of-separation, or small world, phenomenon. Although network nodes like Web pages or people can be aggregated into groups, any given node tends to be a member of several different groups. These groups can be otherwise fairly disparate, but individual nodes spanning them represent short paths between the groups. (See "Groups Key to Network Searches", page 21.)

Some of the same Columbia University researchers have also reproduced Milgram's 1967 six-degrees-of-separation experiment using the Web. The original postal experiment started with 96 messages, 18 of which eventually reached the recipient. The Columbia University researchers experiment prompted 24,163 e-mail volunteers to attempt to reach 1 of 18 target people in 13 countries by forwarding messages to acquaintances. The more thorough experiment confirmed Milgram's broad conclusions, but showed that the original experiment exaggerated the importance of hubs, or nodes that have many connections. (See "Email Updates Six Degrees Theory", page 22.)

Taking short paths

The small-world nature of the Web makes it relatively easy to find information quickly. Researchers are looking for ways to exploit the phenomenon further to speed searching on the Web.

A Cornell computer scientist has combined the small world principal and the concept of fractals to produce an algorithm that makes a network easier to navigate. Fractals are made up of many smaller versions of the same shape and so look the same however much a viewer zooms in or out. (See "Scaled Links Make Nets Navigable", page 24.)

Researchers from Stanford University and Hewlett-Packard's Sandhill Labs have written a search algorithm that leverages the Internet's scale-free structure to get answers to queries more efficiently. Under the scheme, in networks whose nodes keep track of how well-connected their neighbors and neighbors' neighbors are, a node sends a query to the most well-connected node it can find. In networks whose nodes are unaware of how well-connected their neighbors are, a query is passed randomly. The process repeats until a node that can answer the query is found. (See "Search Scheme Treads Lightly", page 24.)

Communities of interest

The surface structure of the Internet is determined by its information content. The simple fact that Web pages tend to link to related pages underlies many of today's search technologies, and is helping researchers developed more sophisticated information retrieval techniques.

This is an important area because the sheer size of the Internet prevents search engines from indexing most of the Web; even combined, the major search engines index less than half of the Web's billions of pages.

A researcher from the University of Iowa has come up with a mathematical model that divides a large network like the Internet into small local Webs based on the idea that Web page authors link to the most popular or important pages in their subject areas. Using this organization Web crawlers could completely traverse a small Web to provide more comprehensive coverage of a given topic, according to the researcher. (See "Webs Within Web Boost Searches", page 25; "Web Pages Cluster by Content Type", page 26.)

Scientists from NEC Corp.'s NEC Research Institute have devised an algorithm that automatically organizes the Web based on the Internet's inherent structure — the way connections grow among pages — rather than text content. The method is designed to solve organization problems that crop up when automatic search results run into ambiguities in text, like identical names that refer to different people. (See "Ties That Bind Boost Searches", page 27.)

And University of Chicago researchers have shown that it is possible to group users across the Web according to common interests based only on their requests for data. The method can be used by anyone, including e-commerce vendors, to target communities of interest. The Chicago researchers are working on using the patterns to design more efficient services for research-sharing environments like Grid computing. Grid computing taps resources around the Web to put together powerful virtual computers that can tackle large scientific problems. (See "Net Scan Finds Like-minded Users", page 29.)

Perceptions

Content influences not only the structure of the Internet but how people use it.

Researchers from Kansas State University have found that the mental models users form as they click through pages on the Internet align more with the way concepts embodied in the pages relate to one another than with how the pages themselves are linked. The researchers showed that common ways of organizing Web information do not follow these content relationships and therefore make for more difficult Web navigation. More closely aligning Web design with mental models would make it easier for people to find what they want on the Web, according to the study. (See "Conceptual Links Trump Hyperlinks", page 30.)

Researchers from the Tel Aviv University and the University of California at Berkeley have showed that English content is likely to continue to dominate the Internet even though use among non-native English speakers is growing at a faster rate than that of native English speakers. English has a first-mover advantage that is common in many kinds of networks, and the advantage is likely to keep English the primary online language, according to the study. (See "English Could Snowball on Net", page 31.)

Weak points

Better searching and browsing are obvious potential benefits of a more complete understanding of the Internet's structure. That knowledge is also critical for preserving the global electronic information commons, because the Internet's scale-free structure harbors vulnerabilities.

Researchers from Clarkson University and Bar-Ilan University in Israel have showed mathematically that when five percent of large hubs are methodically targeted, even large, scale-free networks like the Internet can be broken up into separate islands. (See "Five Percent of Nodes Keep Net Together", page 34.)

The Internet's hubs also make it more vulnerable to viruses and worms. Viruses attach themselves to or replace existing software. Worms, which are less common, are self-contained programs.

A pair of physicists from the University of Notre Dame and the Polytechnic University of Catalonia in Spain have applied condensed-matter physics, which examines the collective behavior of matter, to map the ways viruses traverse the Internet's labyrinth of connections.

The researchers showed that the Internet has become more vulnerable to software viruses in much the same way that human populations that are large and crowded are more likely to fall prey to biological viruses. The Internet is even more vulnerable, however, because connections among computers tend to be more numerous than the human connections that allow biological viruses to spread. The researchers showed that the random inoculation strategy employed for human epidemics does not work on the Internet, but a strategy that targets large hubs does work. (See "Net Inherently Virus Prone", page 35; "Hubs Key to Net Viruses", page 33.)

The nature of networks

Computer viruses bear more than a superficial resemblance to biological viruses. The similar behavior of the two types of viruses points to a fundamental set of properties of all networks. Studying commonalities of and differences between the Internet and other large networks is providing a deeper understanding of networks in general.

A researcher from the Santa Fe Institute has found that social networks are assortative, meaning people who are social gravitate toward others who are social. But the Internet and Web, along with biological networks, are disassortative, meaning highly-connected nodes tend to connect to nodes that have few connections. The practical difference is that in social networks diseases spread easily, but an epidemic is limited in who it can reach, and vaccines can be spread easily as well. The opposite is true in the Internet, Web and biological networks, making them more vulnerable to attack. (See "Social Networks Sturdier Than Net", page 36.)

Scientists from the Wiseman Institute of Science in Israel and Spring Harbor Laboratory have shown that it is possible to categorize networks by finding characteristic recurring motifs, or small local wiring patterns, that occur throughout each type of network. Understanding the function of each motif could help in predicting the behavior of the entire network. (See "Motifs Distinguish Networks", page 37.)

The Internet phenomenon

At a superficial level, the Internet is a mesh network that connects computers and is only remarkable for its size. But that size holds the key to the network's deeper nature.

Because the Internet is too big for any one group or organization to design or control, its structure is governed by tendencies and natural laws. This makes it dynamic, much like a living organism, with change its critical dimension and growth its driving force.

Approaching the Internet as a phenomenon in the scientific sense of the word is allowing researchers to grasp the network's nature, and this, in turn, is opening the door to managing the Internet's growth and exploiting its attributes on a global scale.

Among the goals is fast comprehensive, up-to-date searching. And at the top of the list of important things to figure out about the Internet is its vulnerability to viruses, worms, hackers, terrorists and errant backhoes.

Recent Key Developments

Advances in layout:

- A study showing that the links among Internet routers can be modeled using a nonlinear feedback loop, University of London, February 2004
- A study that shows that the Internet router network has settled into a stable topographical geometrical state (Internet Map Improves Models, page 9)
- A study that shows that the Internet is divided into modules that correspond roughly to countries (Study Reveals Net Parts, page 10)
- An analysis that shows that the evolving hub structure of the Internet puts it at risk (Hubs Increase Net Risk, page 11)
- A study that shows that the physical arrangement of Internet routers follows a fractal pattern (Net Devices Arranged Fractally, page 13)

Advances in links:

- A study that shows that large, well-connected Web sites have more links to each other than to smaller sites (Big Sites Hoard Links, page 14)
- A study that shows that groups of related sites can provide more or less competition for links than the Web as a whole (Odds Not Hopeless for New Web Sites, page 14)
- A study that shows that the faster a segment of the Web grows, the greater the difference between its large and small sites (Faster Growth Heightens Web Class Divide, page 15)
- A study that mapped the link structure of the Gnutella peer-to-peer network, University of Chicago and Argonne National Laboratory, January 2002
- A study that shows that as a Web page becomes older it is less likely to garner new links (Page Age Shapes Web, page 16)
- A study that shows that the rate of appearance and disappearance of links is a critical variable in determining network structure (Disappearing Links Shaped Networks, page 17)
- A study that shows that how frequently links are updated influences the Web's structure (Simulations Sizes up Web Structure, page 18)

Advances in hops:

- A study that shows that naturally-formed networks like the Internet have very few degrees of separation and the number increases very slowly as the network grows (Net Has Few Degrees of Separation, page 19)
- A study that shows that as the Internet grows, the average number of hops between nodes increases only logarithmically (Internet Stays Small World, page 20)
- A study that shows that the small-world nature of networks like the Internet is based on overlapping subsets (Groups Key to Network Searches, page 21)
- An email experiment that verified the six-degrees-of-separation phenomenon (Email Updates Six Degrees Theory, page 22)
- A study that shows that the Internet is a navigable form of small-world network because nodes contain an efficient balance of local and long distance links (Scaled Links Make Nets Navigable, page 24)
- A search algorithm for peer-to-peer networks that takes advantage of the scale-free nature of the Internet (Search Scheme Treads Lightly, page 24)

Advances in content:

- A study that shows Internet search can be improved by exploiting communities of interest (Webs within Webs Boost Searches, page 25)
- A study that shows that Web pages are clustered based on their content (Web Pages Clustered by Content Type, page 26)
- A search algorithm based on Web link structure rather than content (Ties That Bind Boost Searches, page 27)

- A method for identifying communities of interest through Web browsing patterns (Net Scan Finds Like-Minded Users, page 29)
- A study that shows that people form mental models of Web sites based on content rather than link structure (Conceptual Links Trump Hyperlinks, page 30)
- A study that shows that English is likely to remain the dominant language on the Internet because of its head start (English Could Snowball on Net, page 31)

Advances in vulnerabilities:

- An Internet virus inoculation scheme that targets large nodes (Hubs Key to Net Viruses, page 33)
- An analysis that shows that the largest five percent of Internet nodes keep the Net together (Five Percent of Nodes Keep Net Together, page 34)
- An analysis that shows that the Internet's structure makes it prone to viruses (Net Inherently Virus Prone, page 35)

Advances in network comparisons:

- A study that shows that social networks are more resistant to attack than the Internet because highly social people tend to associate with other highly social people (Social Networks Sturdier Than Net, page 36)
- A study that shows that networks of all types can be described by local linking patterns that form recurring motifs (Motifs Distinguish Networks, page 37)
- A study that shows that many types of networks, including the Internet and biological networks, have a scale-free structure (Network Similarities Run Deep, page 38)

Layout

Internet Map Improves Models

By Kimberly Patch, Technology Research News
April 3/10, 2002

As the Internet becomes an increasingly important part of both communications and commerce, it is ever more important to know exactly how it grows.

Scientists from Italy and Spain have analyzed three years worth of daily maps of the Internet's connections in an attempt to better characterize the environment of the real Internet. The study makes a distinction between the Internet, which connects computers around the world in a real network, and software like the Web, which is a virtual network that resides on the Internet.

The researchers used maps of the Internet's connections collected daily by the Cooperative Association for Internet Data Analysis (CAIDA) and the National Laboratory for Applied Network Research (NLANR), which is funded by the U.S. National Science Foundation.

The researchers are looking to closely characterize the Internet because it is an example of a complex network that can be readily analyzed. Their past work was on modeling epidemics and immunization procedures in complex networks like the Internet and networks of human social relationships. "Soon we realized that the deeper analysis, going beyond the simple connectivity properties usually considered, was needed in order to fully characterize the Internet structure," said Romualdo Pastor-Satorras, a visiting professor at the Polytechnic University of Catalonia.

The researchers analyzed Internet growth over time with statistical methods usually used for physics research.

The analysis showed that the Internet grows in specific ways, said Pastor-Satorras. "The Internet can be considered as a spontaneously growing organism. Since there are not global entities regulating the Internet development, it defaults as a self-organized system with high growth rate."

Despite the high growth rate, the Internet has settled into a state whose overall topographical and geometrical properties are stationary in time, said Pastor-Satorras.

"The Internet evolved spontaneously [into] a scale-free network characterized by wild fluctuations in the connectivity properties of the [Internet service providers,]" he said. In scale-free networks, a few nodes, or providers, have many connections to other nodes, while many nodes have few connections.

The analysis also showed that nodes, or computers, on the Internet have settled into well-defined, efficient hierarchies that have to do with how the properties of a node are affected by those of its neighbors. "We find that highly connected nodes are more likely connected to nodes with lower connectivity. This allows us to distinguish different layers of the Internet, or small providers connect to larger providers and so on, following a connectivity and size hierarchy," said Pastor-Satorras.

The hierarchy includes stub domains, which are groups of nodes, or computers that carry traffic only within that domain, or group, and transit domains, which connect different stub domains. The connections inside stub domains are usually short, while the connections among domains are usually long. "The existence of stub and transit domains allows us to identify... a hierarchical structure in the Internet," Pastor-Satorras said.

The growth patterns of the Internet show many short interconnections in stub domains but just a few longer links connecting them to each other, which is "quite [economical] in the sense of the total length of the connections established," said Pastor-Satorras.

The information also allows for the study of properties of Internet service providers like "how the connectivity of providers is related to their age, and the death and replacement events occurring in the growth process," he said.

The hierarchy among nodes, the redundant connections that exist among old nodes, and also the real geographical location of nodes can all influence how the Internet evolves, according to Pastor-Satorras.

The information also allowed the researchers to predict the creation of new connections among providers over time. "This gives information on the forces driving the Internet demand and economical market," said Pastor-Satorras.

The information the researchers have gathered can be used to assess the reliability and effectiveness of computer Internet models used to simulate and test new communications protocols and routing algorithms for the Internet, Pastor-Satorras said. It also may prove useful in developing Internet models, he added.

The researchers are working toward more completely mapping the Internet's characteristics, said Pastor-Satorras. "We would like to have a full characterization of the Internet, including the load of information carried on top of the Internet structure. Since we can consider the Internet as a natural object, we would like to pinpoint the dynamical mechanisms driving the Internet formation and provide a quantitative physical model for Internet growth," he said.

The study sheds new light on some aspects of local connectivity, said Bosiljka Tadic, a physics professor at the Jozef Stefan Institute in Slovenia. The researchers found in the data a correlation regarding who connects to whom, he said. "It appears that many nodes with low connectivity are linked to a few nodes with high connectivity, but in a way that cannot be produced in a generic model," he said.

The research is useful for more realistic modeling of the Internet, said Tadic.

The study also showed that the Internet shares many growth factors with the Web, the virtual network that resides on the Internet. "Apart from several important differences [like] link directions, growth of the Internet... and the Web... are guided in part by the same dynamic rules: growth, attachment [and] rewiring," he said.

The work shows how the topological properties of the Internet are determined by the dynamical properties that govern the network's growth, said Albert-László Barabási, a physics professor at the University of Notre Dame. "We are only at the beginning of our understanding of how this topology emerges in real systems," he said.

The work is important because it combines measurements with simulations to shed light on the interplay between topology and growth and paves the way towards realistic network models, Barabási said.

Pastor-Satorras' research colleagues were Alexei Vásquez of the International School for Advanced Studies in Italy and Alessandro Vespignani of the Abdus Salam International Center for Theoretical Physics in Italy. The research was funded by the International Center for Theoretical Physics (ICTP) and the Spanish Ministry of Science and Technology.

Timeline: Now
Funding: Government
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Dynamical and Correlation Properties of the Internet," posted on the arXiv physics archive at arXiv.org/abs/cond-mat/0105161

TRN

Study Reveals Net's Parts

By Kimberly Patch, Technology Research News
July 2/9, 2003

As the Internet grows, it is becoming increasingly important for software makers and information managers to adapt to the network's basic patterns rather than its current configuration. Researchers studying the workings of the Internet have found several of its structural secrets, but Internet simulations still differ from the real thing.

Researchers from the Nordic Institute for Theoretical Physics in Denmark, Brookhaven National Laboratory, the Niels Bohr Institute in Norway, and the Norwegian University of Science and Technology have uncovered another fundamental Internet attribute — it has an underlying modular structure regulated by the number of sites, or nodes, that link to a given node.

The researchers used a method similar to the algorithm underlying the search engine Google to measure Internet modularity. But rather than the usual method of measuring connected nodes, the researchers focused on links between nodes, mapping out a picture of links linking to links.

They found that the Internet has about 100 modules that correspond roughly to countries, and the farthest points from each other are Russia and U.S. military sites, according to Kasper Astrup Eriksen, who carried out the research at the

Nordic Institute for Theoretical Physics and is now a researcher at Lund University in Sweden.

The work promises to improve the accuracy of Internet simulators, and could help strengthen the Net by pointing out where to reinforce links between weakly-connected modules.

Past research has shown that the Internet is a scale-free network, meaning it has a few well-connected nodes and many nodes with only a few links. "For the Internet the rule is approximately... for every 1,024 nodes with one link, there are 256 nodes with two links, 64 nodes with four links, 16 nodes with eight links, four nodes with 16 links, etcetera," said Eriksen.

In general, scale free networks exhibit preferential attachment, meaning the more links a node already has, the more rapidly it will collect additional links. If a node has two existing links, for example, it is twice as likely to be linked to again than a node with only one existing link.

The Nordic/Brookhaven/Niels Bohr team looked at the structure a little differently, focusing on links between nodes rather than the nodes themselves. Connections can be thought of as being between links rather than nodes, so that a connection to a node with a lot of links is actually a connection to the ends of many links, said Eriksen.

Looking at the structure this way, and keeping in mind that all link ends have the same probability of being picked, at highly connected nodes links "often get a free ride when a new link is connected to one of the other link ends at the same node," he said.

The researchers used a variation on the random walker diffusion method to detect the modularity of the network from the connected links point of view.

Picture a person exploring a network by walking along its links, said Eriksen. "Whenever the walker comes to a node, he picks at random one of the link ends emanating from that node," he said. If you put many walkers on a network and the walkers make decisions independently of each other, they will eventually reach equilibrium — if there are twice as many walkers as links, each link will average at any given time two walkers traveling in opposite directions.

The key to uncovering structural traits of the Internet is studying how this ensemble of walkers slowly reaches equilibrium, said Eriksen.

For example, in a network whose patterns resemble North and South America — with very few links, or roads linking the two parts of the network, or landmasses — a walker starting in North America and turning left and right at random is not very likely to find the road going to South America, said Eriksen.

"What we observed is that first the walkers within North and South America individually come to an equilibrium... and then later on the number of walkers within each country [reaches] its long-term mean," said Eriksen.

Because the walkers reach an equilibrium in a single area, or module first, the method can be used to detect existing modules of the Internet, and can assess the degree of isolation of an individual module, according to Eriksen.

According to the researchers' simulations, the underlying modular structure of the Internet roughly corresponds to individual countries. "We found that the Internet indeed is modular and we identify the large part of this modularity history in the political and geographical divisions in the real world," said Eriksen. The last place the walkers reached equilibrium, for instance, was between Russia and U.S. military sites, he said. "These are thus the... two parts of the network that are most separated from each other."

To carry out the study, the researchers had to adjust some existing algorithms to develop a practical way to run simulation of many walkers, Eriksen said. "Just... running the simulation of random walkers is not the fastest way to calculate the diffusion modes and identify the modules [and is not] feasible for huge networks," said Eriksen.

They found a way to pose the problem so that the time it took the algorithm to calculate roughly doubled, rather than increasing exponentially, every time the network doubled, he said.

The visual results of the simulations were star-like shapes. Straight lines radiating from the center indicated independent modules.

Traditionally, there are two strategies to determining the modularity of the Internet — bottom-up or top-down, said Eriksen. The first scenario groups the most similar nodes into a module. The researchers' work falls into the second scenario, which subdivides the network into modules.

Visualizing modularity is a step toward making a coarse-grain description of the Internet that can be used to better understand its architecture and how and where to improve its connectivity, said Eriksen.

The researchers' method could help make Internet topology generators, or simulators, more accurate, according to Eriksen. "We have devised methods to see if the artificial networks generators are capable of this," he added.

As the Internet grows and changes, it is increasingly important to take its basic patterns into consideration, said Eriksen. "When you devise [software like] routing rules for the Internet, you not only want [it] to do well on today's Internet, but also on the Internet of tomorrow, which at the current speed of development might be quite different," he said. "It is... not good if your algorithms only function efficiently due to... special linkage patterns present today but not tomorrow."

Erikson's research colleagues were Ingve Simonsen of the Nordic Institute for Theoretical Physics (NORDITA), Sergei Maslov of the Brookhaven National Laboratory and Kim Sneppen, now at NORDITA. The work appeared in the April 11, 2003 issue of Physical Review Letters. The research was

funded by NORDITA, Brookhaven National Laboratory and the Norwegian University of Science and Technology.

Timeline: Unknown
Funding: Government
TRN Categories: Internet; Physics
Story Type: News
Related Elements: Technical paper, "Modularity and Extreme Edges of the Internet," posted in the sics Archive at arxiv.org/abs/cond-mat/0212001

TRN

Hubs Increase Net Risk

By Kimberly Patch, Technology Research News
January 1/8, 2003

The Internet has much in common with air travel, according to researchers from Ohio State University. This does not bode well, considering how disruptive storms can be to the airlines.

The commercial Internet has shifted from its original distributed structure toward a hub-and-spoke topology similar to those the airlines use to plot routes, and that shift has made the network more vulnerable, said Tony Grubesic, an Ohio State researcher who is now an assistant professor of geography at the University of Cincinnati.

A handful of cities, including Los Angeles, New York City, Atlanta, Dallas and Chicago, have become central to the Internet and have many more backbone connections than other locations, said Grubesic. These cities are "in effect, acting as hub cities," he said. Chicago, for example, had 23 direct connections to other cities on the AT&T network in the year 2000, versus three for Salt Lake City.

Although the hub-and-spoke topology is cheaper to build, hubs make the network more vulnerable to attack in the same way bad weather in a major hub city can affect flights all over the country, said Grubesic. "Where Internet survivability is concerned, this type of network topology is not a particularly effective one because it forces large volumes of traffic through a handful of cities," he said. "If one of the major points of presence in a city should fail, [for example] the metropolitan area exchange in Dallas, traffic would be disrupted nationwide."

The original topology of the Internet was more distributed, and was designed to withstand failure and provide service under adverse conditions - even a nuclear attack.

As the Internet has grown, however, the competitive nature of the Internet backbone provider industry has caused many providers to shift to the more vulnerable hub-and-spoke system in search of the most economically efficient network topology, according to Grubesic.

The researchers' analysis showed the overall vulnerability of the hub-and-spoke system for 41 network backbone providers. The most susceptible networks have the greatest reliance on hub-and-spoke configurations. The networks most susceptible to disconnection are AT&T, GTE, and Multacom, which would suffer significant performance hits and leave many smaller spoke cities without service with the loss of any one of eight, seven or six of the 14 largest hubs, respectively, according to the analysis.

In contrast, there are 11 network providers that use network topologies that resemble a mesh rather than a hub and spokes; these providers are robust enough to survive the loss of any of the largest hubs. These mesh-like topologies are more expensive to construct, but clearly have advantages where survivability is concerned, according to Grubesic.

To carry out the study, the researchers integrated information about a large set of Internet backbone networks into a geographic information system. "This allowed us to simulate a wide range of Internet disruptions and failures [and] examine... the topological and spatial impacts simultaneously," said Grubesic.

The researchers simulated what would happen if there were a catastrophic failure of the Internet at a hub city or an equally important backbone link. "We simulated the failure of four things: complete loss of a node, or city; loss of a backbone [provider]; loss of a single network node; loss of selected backbone links," said Grubesic.

If an entire hub were knocked out, service to the city in question would be impossible for any backbone. This is a fairly improbable scenario, especially because providers tend to maintain multiple connections in large cities, according to Grubesic. It is a vulnerability, however.

In one portion of the results, the researchers simulated the availability of the network of one provider — Multacom — after a complete node failure.

The city of Washington is the most accessible node on the Multacom network. The researchers showed that if all connections into Tampa failed, Washington would lose access to Tampa plus one other city — Miami. However, if all connections to New York failed, the ramifications for Washington would be much greater; Washington would lose access to New York, Chicago, Denver, San Jose, Portland and Seattle, according to Grubesic.

Worse, the simulation showed that if Atlanta, the most important node on the Multacom backbone, lost all its connections, Multacom communications would cease between Dallas, Los Angeles, Miami and Tampa and 10 other cities each, and between Chicago, Denver, New York, Portland, San Jose, Seattle and Washington and five other cities each.

This scenario is particularly problematic for spoke cities, which rely on the nearest hub.

The second scenario, the loss of a backbone provider, would leave cities serviced by a single provider completely without Internet service. Spoke cities would again be hard hit, according to Grubesic.

The third possibility, failure of a single network node within a city, is a smaller problem. Although this eliminates service

to that node from a single provider, other backbones will remain, allowing traffic to continue, according to Grubesic.

But even a single network node failure would be problematic for spoke cities, because it effectively eliminates the delivery of all traffic destined for the node in question, said Grubesic. The large hubs, and cities served by several providers would do much better because they can reroute traffic.

In the fourth scenario, where select links in a network are severed, isolated nodes would lose service, but nodes that connect to more than one backbone would remain functional.

The methodology can also be applied to other types of networks, including critical infrastructure networks like electric, gas and oil, said Grubesic.

Two of the challenges in carrying out the study were creating code that simultaneously simulated node and link failure for a geographic information system, and developing intuitive ways to interpret the results, Grubesic said.

The researchers' analysis methods can be applied now to the Internet and other types of networks, said Grubesic. "One of our primary goals... was to provide a clear and understandable methodology for estimating the spatial impacts of node link failure for the Internet. This methodology can be revisited, duplicated and perhaps improved by other research teams interested in questions of Internet survivability," he said.

Grubesic's research colleagues were Morton E. O'Kelly and Alan T. Murray. The results are slated to be published in the February, 2003 issue of *Telematics and Informatics*. The research was funded by the National Science Foundation.

Timeline: Now
Funding: Government
TRN Categories: Internet; Computers and Society
Story Type: News
Related Elements: Technical paper, "A Geographic Perspective on Commercial Internet Survivability," *Telematics and Informatics*, February, 2003

TRN

Net Devices Arranged Fractally

By Kimberly Patch, Technology Research News
October 16/23, 2002

Scientists working to make the Internet run more smoothly often rely on simulations of the Net to see how infrastructure tweaks and changes would affect the global network before subjecting the rest of us to live changes.

The trouble is, the Internet's infrastructure is very complicated, and existing simulations tend to oversimplify things. Scientists working to divine the nature of Internet growth are looking to improve these models.

Researchers from the University of Notre Dame have found a clue about network complexity in the physical placement of the Internet's routers, which act as the network's traffic cops. While network models generally place routers at random intervals, in reality routers are physically arranged in a fractal pattern. "The fractal pattern is... the way the routers are placed in space," said Albert-László Barabási, a professor of physics at the University of Notre Dame.

Routers' physical locations are different from the virtual setup of the Net, where links from one router, or node, to the next are dictated by software protocols rather than physical communications lines. The physical router distribution pattern the researchers found is distinct from the virtual link structure, said Barabási. Previous research has already established that the link structure has a fractal pattern. The physical fractal pattern is one more variable that affects the behavior of the network as a whole.

Fractal patterns can be represented by mathematical formulas, and are present in many systems, including fluctuations in the stock market, the distribution of galaxies in the universe, the turbulent flow of liquids, biological growth, and coastlines.

The patterns repeat, and are self-similar, meaning segments of many different-size portions of a pattern look the same as the whole. For instance, the scattered pattern of Internet router locations that shows up on a map of the entire Internet is also discernible on maps of various-size portions of the Internet.

Fractal systems are also nonlinear. Linear systems react in an orderly and predictable way that reflects the magnitude of a stimulus. In contrast, nonlinear systems are much less predictable, and are often very sensitive to small changes.

The researchers found that router density correlates with population density around the world, which is fractal. There are strong, visually evident correlations between router and population density in economically developed areas of the world, according to Barabási. High population density implies a higher demand for Internet services, which results in more routers in those areas.

The pattern of routers' physical distribution comes into play in two ways. The fractal nature of router density is one, said Barabási. In addition, nodes tend not to connect to nodes that are too far away, and this "imposes certain limitations on the network topology," he said.

It takes time and resources to connect Internet routers, and network designers tend to prefer to connect to the closest node that has enough bandwidth, according to Barabási. The costs of physically linking two routers include the technical and administrative costs at the two routers, and the cost and maintenance of the physical line that must run between them. This clearly favors shorter links, according to Barabási.

The fractal pattern of router placement checks with empirical evidence about the structure of the Internet, according to Barabási. This empirical evidence includes the Internet's power-law, or scale-free nature, with a few nodes

having many connections while many nodes have only a few connections.

Combined with recent research into the nature of link bandwidth and traffic, this new information could help network designers anticipate congestion resulting from the Internet's quick, decentralized growth, according to Barabási.

Understanding the impact of router placement on network topology is also important in understanding other types of networks, Barabási said. "The distance dependence is present in social networks as well — you tend to be friends with people at work or who live in your neighborhood, and not with people across the globe," he said. It also shows up in biological networks. "Our brain cells tend to connect to cells that are nearby," said Barabási.

The study is suggestive and worth exploring, said Jon Kleinberg, an associate professor of computer science at Cornell University. It remains to be seen exactly what effect this topology has on performance, he added. "The next step would be to ask how [this] affects the function of the network — the performance of protocols" like those that control routing, he said.

Studies that reveal the structure of networks like the Internet are important because the Net is a growing phenomenon, Kleinberg said. "That it keeps growing and everything keeps working so gracefully as it grows at this unbelievable rate is really because of a huge [amount] of hard work going on beneath the surface — people designing new protocols, simulating them, deploying them," he said. "Working out good models of topology is one component of that."

In a separate study, a group of researchers from Boston University drew similar conclusions about the geographical properties of the Internet.

The Boston University study involved recording the locations of large inventories of Internet routers and links on two occasions two years apart. The study found a quantitative relationship between population density and router density similar to the Notre Dame study, and also showed that router density per person is higher in population centers. The Boston University study also found that 75 to 95 percent of connections between routers strongly relate to the geographical distance between them.

The fractal nature of router distribution could be taken into account in Internet network models within a few months, said Barabási.

Barabási's research colleagues were Soon-Hyung Yook, and Hawoong Jeong. They published the research in the September 30, 2002 issue of the *Proceedings of the National Academy of Sciences*. The research was funded by the National Science Foundation.

The Boston researchers were Anukool Lakhina, John W. Byers, Mark Crovella and Ibrahim Matta. The study was funded by the National Science Foundation.

Timeline: A few months

Funding: Government
TRN Categories: Internet; Physics
Story Type: News
Related Elements: Technical paper, "Modeling the Internet's Large-scale Topology," *the Proceedings of the National Academy of Sciences*, September 30, 2002; Technical paper, "On the Geographic Location of Internet Resources," Boston University computer science department technical report, May 21, 2002, www.cs.bu.edu/techreports/pdf/2002-015-internet-geography.pdf

TRN

Links

Big Sites Hoard Links

Technology Research News, July 2/9, 2003

The Internet is scale-free, meaning it is made up of a few nodes, or servers, that have many links, and many nodes with only a few links. It is also a small-world network — you can get to any node via only a few links among adjoining nodes.

University of London researchers have uncovered another clue about the Internet's structure — the rich-club phenomenon. Large, well-connected nodes have more links to each other than to smaller nodes, and smaller nodes have more links to the larger nodes than to each other.

The researchers found that 27 percent of connections are among the largest five percent of nodes, 60 percent connect the remaining 95 percent to the largest five percent, and only 13 percent of connections are between nodes not in the top five percent.

The findings suggest that the Internet is more dependent on the larger nodes than previously thought, which makes it more vulnerable to a targeted attack, according to the researchers.

The findings could contribute to better strategies for optimizing network traffic flow, network reliability and security, and building network topology simulators; it could be applied to practical systems into three years, according to the researchers.

TRN

Odds Not Hopeless For New Web Sites

By Eric Smalley and Kimberly Patch, Technology Research News
April 17/24, 2002

There is at least a little room at the top, according to a team of NEC researchers who found that the structure of groups of related sites on the World Wide Web is different than that of the Web as a whole.

Recent research has shown that the overall distribution of links on the Web follows a power-law structure, meaning that a small number of large Web sites gain most new links, making them larger still. "This means that an extremely small number of Web pages have the vast majority of inlinks," said David Pennock, a research scientist at NEC Research Institute.

The NEC researchers found that the distribution of inbound links within specific Web communities isn't quite as highly concentrated, however. They used the finding to build a model that accounts for the differing structures of various segments of the Web versus the Web as a whole.

The model can be used as a tool to measure the degree of competitiveness in a network, said Pennock. "This may be important, for example, to e-commerce companies looking to enter a new market niche," he said. The model also applies to natural and social networks like metabolic groups of cells and actor collaborations, he said.

In a rich-get-richer network, nodes that already have many inbound links tend to receive more over time.

"Mathematically, the probability that a node receives another inlink is proportional to the number of inlinks it already has," said Pennock. If a network grows via this preferential attachment process alone, "the rich nodes keep getting richer and the poor nodes can never catch up," he said.

Although it is well established that the rich-get-richer phenomenon applies to the Web as a whole, the research shows that the distribution of inbound links within Web communities is sometimes different from that of the entire Web. Pages within a Web community that contain different numbers of links can have the same probability of receiving a given new link. The researchers refer to this as uniform attachment, in contrast to the preferential attachment of the Web as a whole.

In a uniform-attachment community "more Web pages fare better than would be the case under a pure power-law distribution," said Pennock.

Web communities have a mix of uniform and preferential attachments. For some communities, the percentage of preferential attachment is high and link-poor nodes will have difficulty ever catching up, said Pennock. If the percentage of uniform attachment is high, however, "poor nodes can often — with some luck — get rich, too," he said.

Using the model, the researchers found that the community of e-commerce Web sites selling publications, a category dominated by Amazon.com, is highly competitive and is structured similarly to the Web as a whole. In contrast, the community of e-commerce Web sites selling professional photographers' services is much less competitive, meaning smaller sites have a better chance of gaining links in order to grow.

The research also provides insights into the vulnerability of networks to both accidental failures and malicious attacks, Pennock said. "With more accurate models of different network types, we can begin to understand which are more robust and which are more delicate and prone to disruption," he said.

The researchers' next steps are to measure competition within different Web communities and to apply their model to better understand network fault-tolerance and robustness, said Pennock. They also plan to incorporate some sense of Web page topics because Web pages about the same topic tend to link to one another, Pennock said.

Anyone with access to a search engine and who has the ability to program the researchers' model can now measure the degree of competitiveness within Web communities, said Pennock. Applications of the model in other areas including network fault tolerance and mobile phone networks "are probably about two years off," he said.

Pennock's research colleagues were Gary W. Flake, Steve Lawrence and Eric J. Glover of NEC Research Institute and C. Lee Giles of NEC Research Institute and Pennsylvania State University. They published the research in the April 16, 2002 issue of the *Proceedings of the National Academy of Sciences*. The research was funded by NEC Corporation.

Timeline: Now, 2 years
Funding: Corporate
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Winners Don't Take All: Characterizing the Competition for Links on the Web," *Proceedings of the National Academy of Sciences*, April 16, 2002

TRN

Faster Growth Heightens Web Class Divide

By Kimberly Patch, Technology Research News
June 12/19, 2002

It is clear that the World Wide Web is continuing to grow very quickly, and it is also clear that there is a pattern to the growth. Many research efforts are aimed at figuring out why the Web grows exactly the way it does in order to better plan Web-based efforts like electronic commerce.

Researchers from Korea have put together a mathematical model of the Web that shows that the rate of increase in the number of Web sites influences the pattern of Web growth. One effect is that the faster a segment of the Web grows, the sharper the distinction between its large and small sites — the haves and have-nots of the Internet.

Measurements of the Web have shown that the number of Web sites increases exponentially with time, and that the growth of the Web has a power law, or scale-free structure, with a few Web sites, or nodes, that have many connections to other nodes, and many nodes that have few connections.

To account for the power law behavior, the researchers modeled the interactions among Web sites as an external force acting on any given Web site. Mathematically, this force has both a range and a strength, said Chang-Yong Lee, an assistant professor at Kongju National University in Korea.

Because any Web site can be accessed within a few clicks of a mouse from anywhere else on the Web, there is no spatial limitation to the force range. This is in contrast to the brick and mortar world, where a customer living near a more expensive store is likely to buy an item there rather than at a cheaper but more distant store, said Lee. "For the Web sites on the Internet, however there is no such [geographical] barrier," he said.

Instead, a site's popularity depends on the popularity of all the other Web sites, said Lee. "The force has to be global in the sense that any Web site can act... on any other Web site."

The strength of the force varies in time due to characteristics like Internet topology and changes in traffic, said Lee. Because it is difficult to take into account all the factors that influence the strength of a Web site's popularity, the researchers lumped them together by mapping them all into a time variable.

The model contained one more variable — the growth of the number of Web sites, said Lee. "The number of Web sites at each time step is not fixed, but grows exponentially," he said.

When the researchers used the model to run a simulation of the Web, the simulation reflected most of the important characteristics of the real-world Web, according to Lee. It showed the expected power law structure of links, or visitors to Web sites, and also showed the known relationship between age and Web site popularity, he said. "As time progresses, [a given] Web site will be known to more Web surfers, [and] thus have more popularity."

The model also showed that the strength of the cumulative effects of interactions among sites — the external force — influences the relationship between a Web site's age and its popularity, said Lee. The bigger the strength, the less dependence between age and popularity, he said.

The most surprising result, however, is that the exponent, or curve of the power law structure is directly related to the growth rate of the number of Web sites in the model, meaning the faster the network grows, the bigger the difference between the large and small nodes, said Lee.

This means that different segments of the Web that have different growth rates also have different power law curves. For example, when numbers representing the Web as a whole and the .edu portion of the Web are plotted on a graph, with one axis showing the number of sites and the other axis showing site size, the steepness of the curve connecting these numbers is different for the two categories.

Within the researchers' model, this variation can be explained by looking at different growth rates of the number of Web sites for each category, said Lee. "The larger the growth rate, the bigger the exponent," he said. In graphical terms, the power law curve gets steeper as the growth rate increases, making the overall growth rate curve for the Web steeper than the curve representing just the .edu domains.

The Web is essentially a competitive, complex system that is always changing as the number of Web sites increases exponentially, according to Lee. The global interaction of the system and its dynamic nature are the two main ingredients that determine its structural characteristics, he said.

The researchers' model may capture some properties of the .edu domain, but the work doesn't necessarily carry over to the general network, said Bosiljka Tadic, a physics professor at the Jozef Stefan Institute in Slovenia.

The research results are potentially interesting for the .edu domain, which has a power law distribution close to that of the model, said Tadic. The model could be used to estimate the "projected number of visits given that the trends in growth of the number of nodes are known," he said.

Lee's research colleague was Seungwhan Kim, of the Korean Electronics and Telecommunications Research Institute. They published the research in the March, 2002 issue of the journal Physical Review E. The research was funded by The Korea Science and Engineering Foundation (KOSEF).

Timeline: < 5 years
Funding: Government
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Dynamic Model for the Popularity of Websites," Physical Review E, March, 2002

TRN

Page Age Shapes Web

By Kimberly Patch, Technology Research News
November 21, 2001

As it turns out, networks like the Web have something in common with humans — how they are structured has a lot to do with aging.

Two researchers from the Niels Bohr Institute in Denmark have found they can more accurately model networks like the Web by using a sort of memory that takes into account the potential for waning popularity as a Web page gets old. They also found that this age variable plays a part in the stability of real world networks.

This is a step forward from earlier models, which tend to represent popularity simply as the number of links a node has accumulated.

The older a Web page is the more likely it is to go out of date and become less popular, said Konstantin Klemm, a graduate student at the Institute. "After an exciting Web page

has been launched, many other pages link to it. However, sooner or later the novelty and consequently the popularity of this page decreases. Interest shifts to pages released more recently, which are now the ones that rapidly increase the number of incoming links."

The age variable essentially introduces time into the network structure model, said Klemm. This makes for more accurate simulations of real world networks, which can potentially help in predicting how networks like the Web and wireless telecommunications will grow, he said.

It also stabilizes the network, said Klemm. "This gives insight into which growth mechanisms lead to stable structures, which can be very useful in devising strategies for expanding telecommunications networks," he said.

The model gives each node a certain amount of staying power based on how novel its contents are. "Each node in the network is assigned an extra degree of freedom containing the ability to attract links — [its] popularity," said Klemm. When a node is created it is linked to nodes that are popular at the time. It then receives links from nodes created subsequently, he said. "This continues until eventually the node under consideration loses its popularity."

The model mimics the appearance of new pages on the Web, which have links to popular pages instead of outdated, forgotten ones, Klemm said. "A Web page released at the beginning of the year 2002 will most likely contain many links to pages of the year 2001, [and] fewer references to 2000," he said. The researchers made the assumption that the more links point to a given Web page, the less likely it will be forgotten, said Klemm.

When the researchers tested their model, they found it more closely maped real-world networks than previous ones. Many networks, including the Web, have several structural traits in common.

Scale-free networks have a few nodes with many links and many nodes with only a few links.

In small-world, or six-degrees-of-separation networks, every node can reach every other node through a relatively small number of hops, or jumps from one node to another. This small-world trait emerges when nodes in a network become clustered into groups, and is also true of networks of people, who cluster according to dynamics like where they live, work, or what interests they have.

Many network models show how either the small-world or the scale-free structure arises. Both structures emerge from the researchers' simulations, according to Klemm. "Our work connects the concepts of growing scale-free networks and the small-world effect. In our model the crucial properties of the network emerge from one simple continuous growth process." Clustering develops in the model due to the age variable, he said. "This... locality in time... causes clustering," he added.

The work is "an important step ahead in network modeling," said Reka Albert, a research associate at the University of Minnesota. There are two major directions in network modeling: small-world networks and scale-free networks, she said. The Niels Bohr Institute model "unites the large clustering coefficient of small-world network models and the... distribution of scale-free network models."

The model is useful because it closely maps real world networks, many of which have at least some amount of node memory, said Albert. "In citation networks there's a clear tendency to refer to more recent papers. In collaboration networks... new collaborations are restricted to the active people."

The Web is also affected by some degree of node memory, but the situation is not as clear-cut, Albert said. This is because it not only contains both pages with unchanged content that become outdated and forgotten, but also well-maintained pages that receive new links independent of their creation time, she said.

The next step in the research is to study dynamics like communication between nodes and propagation of diseases in their network models, said Klemm. The research should eventually point the way toward better search strategies on the Web. The idea could be applied practically within five years, he said.

The research was funded by The Danish National Research Council.

Timeline: < 5 years
Funding: Government
TRN Categories: Networking; Internet
Story Type: News
Related Elements: Technical paper, "Highly Clustered Scale-Free Networks," posted in the arXiv physics archive atarXiv.org/abs/cond-mat/0107606

TRN

Disappearing Links Shape Networks

By Kimberly Patch, Technology Research News
November 14, 2001

It is fairly obvious that networks are a common ingredient in social circles, biological processes and computer communications. What is much less apparent is exactly how the different aspects of these networked systems interact to direct network growth.

Researchers are delving deeply into network dynamics to try to tease out what makes a difference in the growth of things like working relationships among actors, connections among biological processes in cells, and links among Web pages on the Internet.

A better understanding of these relationships promises to help the Internet grow more smoothly and may make mobile networks like those of cellphones easier to manage.

One common ingredient in many large networks is a scale-free, or power-law structure. In scale-free networks, a few nodes have a lot of connections to other nodes, and many nodes have only a few connections each. Previous research has shown that this structure can be caused by a sort of perpetual rich-get-richer dynamic that says the larger a node is, the more likely it is to attract links.

A theoretical physicist from St. Petersburg State University in Russia has found that this common network structure may also be maintained by a different dynamic. Her disappearing link model shows that under certain conditions the rate of appearance and disappearance of links in a network may also cause a scale-free structure.

The research also shows that the rate of appearance and disappearance of links is probably more important to the dynamics of the network then the size of the network.

In real-world networks like human social circles and the Internet, the dynamic of links disappearing is common, but this dynamic is often absent in research studies of networks, said Olga Kirillova. Investigations of how significant "the maximum number of possible arising and disappearing links is [are] practically absent," she said.

Events like movie actors fading away and Web pages doing the same may be important to network structure, according to Kirillova. In real communication networks there are phenomena such as species extinction, aging and death, she said. These changes don't just simply subtract a node from the system; they also change the network's structure of interactions, she said.

Kirillova's research showed that when the number of links appearing and disappearing was set at about 0.5 percent of the network, the network was pushed toward a scale-free structure. The rates of link appearance and disappearance in both the Internet and in the network of scientific paper citations fall close to this number, according to Kirillova.

The model is a valuable one that may bear on several types of real world networks, said Bosiljka Tadic, a theoretical research scientist at the Jozef Stefan Institute in Slovenia. "In particular, it may be useful for closed communities and on a relatively small time scale. For instance, [biological] food chains and commercial supply networks... may be sensitive to fluctuations of links. Cutting a link or adding new link may trigger a cascade of link updates," he said.

A big question is why so many networks have a scale free structure, "because we seem to see it everywhere," said Jon Kleinberg, an associate professor of computer science at Cornell University. "The question is what is the... basic mechanism at work that's causing all these networks to have this power-law structure," he said.

Kirillova's research says if you "correlate the appearance and disappearance just right you get this power-law behavior even if you don't have a rich-get-richer kind of process," said Kleinberg. "This is yet another way to see power-law" type networks arising.

In the end, the problem of network structure is getting "more challenging because it isn't that there is somehow a single explanation" of the scale free structure, said Kleinberg. "It's completely conceivable that they are arising in different situations for different reasons." This, in turn, raises an important issue, he said. "When we see a power-law [structure], how do we decide... which model... is really the best approximation?"

Kirillova's model may have a significant role in exploring the behavior of wireless networks, whose links appear and disappear fairly quickly, Klineberg added.

Network models may also provide insight into biological evolution, said Kirillova. One of the most important aspects of evolution is that useful structures like limbs and organs emerge through slow improvements that are sparked by random genetic changes and limited by the laws of physics. Understanding the emergent, dynamic structure of networks, which also harbor local rules that govern changes, may eventually help us better understand biological evolutionary processes, she said.

Kirillova published the research in the August 6, 2001 issue of *Physical Review Letters*. The research was funded by St. Petersburg State University.

Timeline: Now
Funding: University
TRN Categories: Networking; Internet
Story Type: News
Related Elements: Technical paper, "Communication Networks with an Emergent Dynamical Structure," *Physical Review Letters*, August 6, 2001

TRN

Simulation Sizes Up Web Structure

By Ted Smalley Bowen, Technology Research News
January 17, 2001

With every hyperlink added to a Web page the shape of the World Wide Web changes, but the problem of determining that shape is trickier than just estimating the number of links and mapping where they go.

Large, complex, evolving networks have proven difficult to model statistically, and the Web is a particularly challenging case.

A model and simulation that looks at how frequently links are updated and at the relationship between outgoing and incoming links promises to give a clearer picture of the Web's underlying structure.

The model, developed by Slovenian researcher Bosiljka Tadic, suggests that the shape of the Web is largely determined by how and how often Web pages are updated, which in turn is determined by the biases and policies of the people who

update the pages. The research also showed that the underlying structure of the Web is still maturing.

"Understanding the underlying dynamic rules that govern the Web may help to design more efficient algorithms for discovering [what makes the Web change] and predicting [its] evolution," said Tadic, a research scientist and physics professor at the Jozef Stefan Institute in Slovenia.

Tadic's 5-million-node simulation represents the Web as a directed graph, showing nodes, or Web pages, connected by arcs, or links. It compares link distribution and the size and depth of nodes with statistical snap-shots of the actual Web. The simulation correlates fairly closely with actual Web topology and growth data, said Tadic.

Although the Web has many similarities with other large, complex networks, one of its differences is how quickly the links between nodes change. In contrast, links in many other social networks change slowly or not at all, Tadic said.

The model assumes that because Web links are constantly updated, the network is being rearranged at the same pace at which it grows.

On the Web, outgoing and incoming links are both hierarchical, but show different patterns and rates of growth, Tadic said. A link is classified according to its relationship to a particular page: clicking on a hyperlink on a page generates an outgoing link, while a visitor arriving at the page via an outside link generates an incoming link.

The key to Tadic's simulation is that it takes into consideration this relationship between the patterns of the two types of links, assuming that separate but related rules govern the growth of inbound and outbound Web links. "In the Web, the incoming links are driven by the dynamics of outgoing links," Tadic said.

Statistical models of the Web generally do not factor in these critical elements of how links to and from sites change and how they might be related, according to Tadic.

The model also draws from recent research into network node sizes. It assumes that the most active Web masters create the bulk of the network's links, that the most popular sites have by far the most links, and that there are many, many nodes with just a few links.

Tadic is applying to the Web general ideas developed in the last year or so about network modeling, said Albert-László Barabási, associate professor of physics at Notre Dame University.

"The basic idea is that if we understand the Web better, then we can design better search engines. Any tool that works on the Web would work better if we start with a better understanding of how the Web develops and what is its topology. In that sense, this paper is very valuable," he said.

Further research into this area might address the Web's clustering properties, and examine the Web as a grouping of several weakly linked networks, Tadic said.

The model could also apply to other types of complex evolving networks, he said.

Tadic's research has been accepted for publication in the journal *Physica A: Statistical Mechanics and its Applications*. The research was funded by the Ministry of Science and Technology of the Republic of Slovenia.

Timeline: Now
Funding: Government
TRN Categories: Internet, Computers and Society
Story Type: News
Related Elements: Technical paper, "Dynamics of Directed Graphs: the World-wide Web," accepted for publication in *Physica A: Statistical Mechanics and its Applications*. It is posted, along with related papers at phobos.ijs.si/~tadic/paperslist.html

TRN

Hops
Net Has Few Degrees of Separation

Technology Research News, March 12/19, 2003

Researchers from Bar-Ilan University in Israel have found that the average number of connections needed to get from one point to another in real-world networks like the Internet and social networks is smaller than the number needed for randomly-connected networks.

Accounting for the difference could improve traffic flow and even provide better virus protection.

The finding has to do with the small world concept, which says that any two people in the United States are connected by less than six degrees of separation.

The researchers found that in naturally-formed networks — like groups of people or the Internet — the degrees of separation are fewer than in a randomly-connected network model, and this number increases extremely slowly as a network grows.

Randomly-connected network simulations are often used in designing Internet tools. The researchers' work can be used to design tools that route traffic more efficiently, improve searches, and better immunize networks against viruses. It could also be used to design networks that have shorter paths between points.

The method can be used today to develop algorithms to improve the workings of networks, according to the researchers. The work appeared in the February 7, 2003 issue of *Physical Review Letters*.

TRN

Internet Stays Small World

By Kimberly Patch, Technology Research News
September 12/19, 2001

The challenge to keeping information flowing smoothly over the Internet is being able to identify what affects the flow of data packets and how this will change as the Web grows. Only then can you plot the most efficient way to direct Net traffic.

This is harder than it sounds. The way traffic is routed on the Internet today is not as precise as it could be because live tests on such a large network are difficult and expensive, and it is difficult to simulate all the variables of networks as complicated as the Internet.

Researchers from Stanford University, however, have found a way to more closely simulate the way information flows over these types of networks. And in looking at the simulations they found good news — there are potentially more efficient ways to route traffic over the Net, especially as it grows larger. The simulations also shed some new light on the principles behind the classic six-degrees-of-separation experiment.

The Stanford simulation algorithm took into account the scale-free, or power law nature of the Web; its small-world, or clustering nature; and a trait known as short path links. In scale-free, or power-law networks, a few nodes have many links to other nodes, while many nodes have only a few links. In clustering networks, the relationships among nodes are not randomly distributed, but are grouped. Short path links means there are some very short paths sprinkled throughout the network that may directly link one group to another.

Previous algorithms have allowed researchers to simulate scale-free networks and small-world networks with short path links, but not networks that harbor all three traits. There are many natural networks that exhibit all three traits, including social networks that describe relationships among groups of people, and metabolic networks that describe things like the substances a cell uses in life processes.

The researchers found that as networks that harbor all three traits grow very large, they become very efficient in the number of steps, or hops a data packet takes to get from one node to another node. "Surprisingly it took a far shorter number of steps for a scale-free small world than for a traditional small world," said Amit Ram Puniyani, a physics graduate student at Stanford University.

While the average number of hops grew from 10 for a 10,000-node network to 200 for a 100-million-node network in networks that were scale-free or small-world, the number of average hops grew much more slowly in networks that were both scale-free and small-world, said Puniyani. The number of steps grew "logarithmically with the size of the network, which means that for 10,000 nodes you need five steps, [but] for 100 million the number grew only to 6.5," he said.

This relationship may also explain the reason behind the six degrees of separation found by the classic Milgram experiment, Puniyani said. In the mid-60s, when the population of the United States numbered around 190 million, social psychologist Stanley Milgram asked people living in Omaha, Nebraska to pass messages to a target person in Boston through a chain of acquaintances. People receiving the letters sent them to an acquaintance who was most likely to lead to the target. "Milgram found that it took an average of six steps for the letters to reach the target. We believe we have explained this mysterious number," Puniyani said.

The models for networks that were scale-free or small-world, but not both, predicted an average of more than 200 steps for a letter to find the target person in a network as large as the relationships among all the people in the U.S. The researchers' logarithmic growth curve, however, more closely matches Milgram's empirical evidence that an average of only six hops was needed for information to get from one node to another in a 190-million person network.

The work is extremely relevant and the method appears particularly efficient and fast, said Alessandro Vespignani, a research scientist at Abdus Salam International Centre for Theoretical Physics in Italy. "The search algorithm is especially devised for scale-free networks and uses their wildly fluctuating connectivity to speed up the search time. This is both novel and interesting," he said.

The significance of the research is it shows how data can be sent from one place to another in a relatively short period of time without having to know the whole path the data must travel, Vespignani said. In large networks like the World Wide Web, the path a given piece of data must traverse to reach its destination can be incredibly complex, he said.

The researchers algorithm has many potential uses, Vespignani said. "The algorithm could be useful in a huge number of tech applications, especially in [optimizing] addressing and searching times on the Internet and the World Wide Web. In principle this could lead to new routing procedures for sending data to Internet addresses," he said.

The researchers are working on making the model more realistic by adding differential clustering, said Puniyani. Differential clustering would add to the model the complexity that "the probability of linking between two nodes decreases as the distance along the ring [of neighbors] increases," he said.

In a social network, for example, this would take into account the lesser probability that an American has acquaintances in Australia than in America, and the increasing probabilities that a person knows someone who lives in the same city, works in the same office or lives in the same neighborhood. "This would make the whole thing much more realistic and appropriate for simulations," said Puniyani.

The existing algorithm can be used for evaluating the efficiency of traffic routing algorithms and for doing more realistic modeling in areas like epidemiology and economics, Puniyani said.

Puniyani's research colleagues were Rajan M. Lukose and Bernardo A. Huberman of Hewlett-Packard's HP Labs. The research was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Networking; Internet
Story Type: News
Related Elements: Technical paper, "Intentional Walks on Scale Free Small Worlds," posted on the Los Alamos physics A rXiv at arXiv.org/abs/cond-mat/0107212

TRN

Groups Key to Network Searches

By Kimberly Patch, Technology Research News
May 29/June 5, 2002

Sociologists and marketers alike recognize that links between people follow patterns that can be exploited to more clearly understand group behavior.

One tantalizing clue to the way very large groups of people are connected is the tidy 1967 result of sociologist Stanley Milgram's postal experiment. The six-degrees-of-separation cliche was spawned when Milgram found that it took an average of only six exchanges, or hops, between people and their acquaintances for a letter to find its way from a person in Omaha, Nebraska to a Boston recipient the original sender did not know.

It's taken much longer for scientists to tease out a theory that explains Milgram's empirical evidence.

A group of researchers from Columbia University have constructed a mathematical model that explains just how this can be. The model promises to provide insights into social behavior and also shed light on the structure of other networks, like the World Wide Web. The relationship between two people who know each other is analogous to a link between Web pages. The work could lead to better search techniques for the Web.

Groups are the crux of the matter, according to Duncan Watts, an associate professor of sociology at Columbia University. "We all belong to groups, and the set of groups each of us belongs to is one way to characterize us."

Groups are responsible for determining who we meet and helping us measure how similar we are to others, said Watts. "So when I show you a description of someone and you think 'I am nothing like this person' you're really thinking 'I don't belong to any of the social groups that this person belongs to, therefore I'm not likely to run into [him].'" Someone who belongs to a country club in Bel Air, for instance, is unlikely to be in the same group as a Georgia farmer.

What makes the six-degrees-of-separation, or small-world, phenomenon possible is that although we tend to aggregate into groups, any given person tends to be a member of several groups, said Watts. "This is where the trick is," he said. Although we tend to associate with people who are like us, we have more than one way of assessing these similarities, Watts said. "For instance, you're close to the people you work with. And you're close to the people you went to college with. But they're not necessarily all that close to each other."

Because of this, individuals can span very different groups, or social dimensions, said Watts. Take, for instance, three people: A, B, and C. A can be close to B in a group defined by geography, and B can be close to C in occupation, but A and C may perceive each other as far apart.

To get from one person to any other, a message can be directed through these groups to find its target relatively quickly, said Watts. "As long as A knows that B is more like C than A... all A needs to do is pass the message to B and rely on B having better information. B then makes use of her other dimension to direct the message," he said.

Previous research pointed out that if Milgram's results were true, these types of short paths must not only exist in social networks, but people must be able to find them without much information about the world, said Watts. "Our contribution has been to show how this can be done in a way that is sociologically plausible," he said.

Surprisingly, the model showed that people don't have to belong to very many groups for the small-world phenomenon to kick in, said Watts. The optimal performance of a social network occurs when individuals are members of an average of only two or three groups, which is the number people actually tend to be in, he said. "We expected there to be a trade-off between too few and too many social dimensions, but we didn't expect the optimal number to be so low," he said.

Ultimately, there is more to a network than the pattern of connections between people or Web pages, said Watts. Network nodes like people and Web pages "have classifiable properties that predate the network structure," said Watts.

A full understanding of the structure of a network requires an understanding of this social structure, which is, after all, what brought about the network's connections, Watts said. "You can't understand the network structure without first understanding social structure. They're related, but they're not the same thing," he said.

The model could eventually improve the algorithms used for searching computer networks like the Web, said Watts. This is a case of observing people's behavior, then teaching it computers. "We're... reverse-engineering an empirically-observed capability that people in social networks seem to possess," and using it to solve problems in computer networks, he said.

This natural social model is different from the traditional computer science approach of building complicated search software that operates over a relatively simple network structure, said Watts. The structure of the social network is more complicated, and requires only simple search strategies. "The capabilities... are not due to people using particularly sophisticated [methods] for conducting searches. Rather, the bulk of the work is done for them by the network, which is built in just such a way that even a simple search procedure works," he said.

The model may also have practical applications for sociological problems. It could lead to ways to improve people's access to information through their social networks, said Watts.

Understanding how messages and ideas travel in social networks is an open problem in both sociology and marketing, said Albert-Laszlo Barabasi, a physics professor at the University of Notre Dame and author of the recent book Linked: The New Science of Networks. Structure is easier to analyze in networks like the World Wide Web than in social networks because search engines can map out how pages are connected to each other, he said. "We're missing such tools for... society," he said.

The researchers' searchable model arranges societies' links into a hierarchical topology based on shared geographical habits and interests. "This is an interesting hypothesis, which indeed allows them to explain certain features of how messages travel," said Barabasi.

The work may also provide insights into the Web, Barabasi said. "Such shared-interest-based local organization could be present within the Web as well."

In addition, the Web could help quantitatively prove that this type of hierarchy is present in networks, said Barabasi. "One important step... needs to be taken," he said. The researchers hypothesis should be tested on the link-based topology of the Web, he said.

The researchers are currently turning their model by gathering more empirical evidence. They are also planning to look into what this new network knowledge means for network properties other than searching, said Watts. The model could be used for practical purposes like improving Web searches within two years, he added.

Watts' research colleagues were Peter Sheridan Dodds of Columbia University and Mark E. J. Newman of the Santa Fe Institute. They published the research in the May 17, 2002 issue of the journal Science. The research was funded by the National Science Foundation (NSF), Intel Corporation, and Columbia University.

Timeline: 2 years
Funding: Corporate, Government, University
TRN Categories: Internet
Story Type: News

Related Elements: Technical paper, "Identity and Search in Social Networks," Science, May 17, 2002; Small World Research Project: smallworld.sociology.columbia.edu

TRN

Email Updates Six Degrees Theory

By Kimberly Patch, Technology Research News
August 27/September 3, 2003

The world has known about the small-world phenomenon since sociologist Stanley Milgram's 1967 study found that it took, on average, six exchanges among acquaintances to get a letter from a random correspondent in Omaha, Nebraska to a Boston recipient identified only by a brief description.

But the experiment, which started with 96 messages, 18 of which eventually reached the recipient, has been criticized for not being thorough.

Columbia University researchers have filled in the blanks by carrying out a larger, more detailed experiment over the Internet. The results match many of the broad conclusions of Milgram's work, but show that Milgram's conclusion about the importance of hubs — people who have many connections — may be off, at least in regards to social networks.

The results could contribute to better knowledge bases and peer-to-peer network design.

The researchers' global social search experiment, posted on the Internet, prompted 24,163 email volunteers to attempt to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. "People were given the description of the target individual and asked to send an email to a contact of theirs who they thought was in some way closer to the target," said Peter Sheridan Dodds, an associate research scientist at Columbia University.

Recipients of these messages were instructed to do the same until the message reached the target individual. The researchers also asked participants for demographic data and their reasons for choosing the contacts they did, said Dodds.

The experiment generated a total of 61,168 emails that ran through 166 countries; 384 of the original 24,163 reached their target. The researchers concluded that individual apathy or disinclination to participate was the major reason for broken chains. When they questioned senders who did not forward their messages after one week, only 0.3 percent said they could not think of an appropriate recipient, according to Dodds.

The experiment confirmed that, in social searches, a message initiated by a random person reaches its destination in a median of five to seven steps, depending on the separation of source and target, said Dodds. "People can find other people regardless of how distant they are," he said. "Search in large-scale networks is a very difficult problem, and yet people working collectively are able to do quite well."

With a greater number of successful searches to analyze, however, they found something surprising: the primary avenues from source to target are not the highly connected social hubs that Milgram's experiments pointed to. "Successful chains were... far less likely to use hubs than unsuccessful chains," said Dodds. "Hubs — people with many friends — don't seem to be so important for this kind of social search."

Participants in successful chains were less likely to send messages to hubs — 1.6 percent versus 8.2 percent — than those in incomplete chains.

The participants' answers also reflected this. They rarely chose a sender based on the number of friends that person had, according to Dodds. The main reasons for choosing the next person were geography- and work-related," he said. These two categories accounted for at least half of all choices, with geography dominating in the early stages of the chain.

When the researchers compared successful chains to those that did not reach their target, they found that successful chains involved more professional ties — 33.9 percent versus 13.2 percent — and fewer familial relationships — 59.8 percent versus 83.4 percent.

The experiments also showed that the success of a search is highly dependent on individual incentives, said Dodds. This is because a small change in the attrition rate — the probability that people don't send a message on — leads to a substantial change in the number of chains getting through, said Dodds.

One of the researchers' targets, a U.S. professor, received many more completed chains than any of the other 10 targets reached. This was probably because the professor appeared reachable, which makes sense because the participants were largely college-educated and 50 percent lived in the U.S., said Dodds. "We interpret this as meaning that people's perceptions greatly affect their chances of success," he said. "In a nutshell, if you think the world is small, it is."

The study is a confirmation of the six degrees of separation in social networks, but also debunks some ideas associated with the six degrees that have entered the popular culture, said Stephen Strogatz, a professor of applied mathematics at Cornell University. "Milgram didn't really have enough participants to figure out what kind of methods people were using" to reach the target, he said.

The new study shows social networks don't depend on super-connected people, or hubs, said Strogatz. "There was no evidence for hubs in this study, and yet people were still able to get the messages to the targets in the successful chains in five to seven steps," he said.

Although there are many similarities between social networks and virtual networks like the Internet, it makes sense that hubs may not be as prominent in the real world, said Strogatz. When hubs are virtual, like on the Internet, "there's no physical or economic cost to having many people point to your Web page," he said. In the real world, however, network hubs contain costs — maintaining a rolodex of 100,000 people, for instance, takes time, and maintaining a powerplant with many transmission lines costs money.

In the Milgram study, all successful chains went through one person — a well-connected tailor. The Columbia study did not show this funnel effect, however, said Strogatz. It showed, rather, "that there are a lot of roads to Rome," he said.

The new research also shows that the key to good social searches is weaker friends, or more distant acquaintances, said Strogatz. This makes sense — closer friends are less useful in this case because people who know each other well tend to have the same friends, he said.

"This is good work," said Jim Moody, an assistant professor of sociology at Ohio State University. "By understanding the structure of email communication networks we might be able to better design tools for spreading information or stopping virus flow," he said.

The weak-tie findings are consistent with research about how people use their acquaintances to find information when they're looking for work, said Moody.

More work on the structure of close ties is needed, Moody said. "The weak-tie findings... will eventually be quite important for information flow," he said. "For virus flow through email, the key rests on the structure of close, trusted email contact — we won't open attachments from people we don't know well."

The next step in the research is an experiment currently on the researchers' Web site that is designed to gain more information about how the small-world phenomenon works, said Dodds. "People can now send more than one email regarding any given target," he said. Also, "we've altered the questions we ask about why people choose the people they do, and extended the descriptions of the targets."

The current results could be used now to improve databases and networks, according to Dodds. "This... could be useful in the design of knowledge bases or in the construction of peer-to-peer networks," he said.

The researchers are also working to model a range of social and economic problems including the spread of contagion agents like diseases or fads, the evolution of cooperation, and the structure of modern organizations, said Dodds.

Dodds's research colleagues were Roby Muhamad and Duncan J. Watts. The work appeared in the August 8, 2003 issue of *Science*. The research was funded by the National Science Foundation (NSF), the James F. McDowell Foundation and the Office of Naval Research (ONR).

Timeline: Now
Funding: Government, Institution
TRN Categories: Internet; Applied Technology; Computers and Society
Story Type: News
Related Elements: Technical paper, "An Experimental Study of Search in Global Social Networks," *Science*, ugust 8, 2003;

Researcher's book, "Six Degrees, The Science of a Connected Age," by Duncan Watts; Current small-world phenomenon experiment Web site: smallworld.columbia.edu

TRN

Scaled Links Make Nets Navigable

By Eric Smalley, Technology Research News
September 6, 2000

Actor Kevin Bacon in a fractal T-shirt makes a pretty good symbol for a mathematical principle that could yield better routing algorithms and search engines. The principle, discovered by Cornell computer scientist Jon Kleinberg, describes the most efficient structure for networks like the World Wide Web.

Kevin Bacon is the poster boy for the concept that anyone can reach anyone else through six or fewer someone-who-knows-someone connections. And the fundamental principle behind fractals is that the geometries of a particular image remain the same however much a viewer zooms in or out. This is because each shape is made up of many smaller versions of the same shape.

Combine the notion of six degrees of separation with the same-at-every-scale property of fractal geometry and you have in a nutshell Kleinberg's algorithm for making a network easier to navigate.

People don't need a bird's eye view of a network to navigate it efficiently. We wanted to understand what about networks makes that possible, Kleinberg said.

Kleinberg found the answer in the structure of network connections. Many networks have short chains, which means they have relatively few links between any two points. But some short chain networks are easier to navigate than others. It's harder to find the short chains in networks where individual nodes have substantially more local links than long distance links or vice versa.

Put in terms of a social network, the most efficient structure for navigating a network is one in which individuals have as many friends in their counties as in their towns minus the friends in the towns, as many friends in their states as in their counties minus the friends in the counties, and as many friends in the country as in their states minus the friends in the states, Kleinberg said.

"These local navigation algorithms seem to work best when the network is equally rich in links at all of these length scales. This is a network with a sort of built-in gradient that funnels you toward targets. No one individual knows the short chains. But if they simply start forwarding [a message] in the right direction then somehow the structure of the network actually works to funnel it in on the target," he said.

In addition to figuring out how this network principle can be used to improve the Internet, Kleinberg is relating this information to how people use the Web. "As people close in on good information, are they actually making use of the cues that are available?" he said.

Kleinberg's work was funded by the National Science Foundation, the Office of Naval Research, and The David and Lucile Packard Foundation.

Timeline: Now
Funding: Government, Private
TRN Categories: Networking
Story Type: News
Related Elements: None

TRN

Search Scheme Treads Lightly

By Kimberly Patch, Technology Research News
June 6, 2001

In social circles, people usually know who would know who would know. In other words, we are good at knowing where to start asking around for information.

Researchers from Stanford University and Hewlett-Packard's Sand Hill labs are applying the concept of knowing whom to ask first in order to design more efficient search strategies for the Internet.

The work is based on recent research that shows that both social circles and the Internet are scale free, or power-law networks, which harbor a few nodes or people with many, many connections, and lots of nodes that have only one or two links.

The Stanford and Sand Hill researchers wrote a search algorithm that takes advantage of this structure in order to get answers to queries much more efficiently. "In a way, the strategy is already used. People naturally ask people who they consider well-connected to put them in touch with someone who knows about a particular topic," said Lada Adamic, one of the Stanford researchers and a consultant at Xerox Palo Alto Research Center (PARC).

The algorithm works in one of two ways:

In a network where each node keeps track of how well connected its neighbors and its neighbors' neighbors are, a node sends a query to the most well-connected node it knows of. If this node cannot answer the query, it sends it on to the most well-connected node it knows of. The process repeats until a node that can answer the query is found.

In a network where the nodes are unaware of how well connected their neighbors are, the query is passed on randomly. Though less efficient, the method scales reasonably well simply because large nodes have a lot of paths leading to them and so tend to attract queries.

The algorithm is useful in finding information in peer-to-peer networks distributed over the Internet, like Freenet and Gnutella. Nodes in peer-to-peer networks communicate with

each other without using a central server to direct the traffic. Napster, for example, uses a central server.

The decentralized nature of peer-to-peer networks makes them less vulnerable to attack, because in order to shut down the network all of its member nodes must be shut down. The lack of a central server, however, makes them considerably less efficient in fielding requests for information.

This is because, lacking central direction, each node passes a request to its neighbors, which then pass the request to their neighbors. Eventually the request makes its way to the server harboring the information.

The trouble is, lots of nodes are involved in every request for information. "In the current Gnutella search algorithm, every node passes the query on to every one of its neighbors, which means that the number of nodes possessing the query grows exponentially with the number of steps," said Adamic. It takes an average of 4.3 steps to find 50 percent of the nodes in an 800 node peer-to-peer network, but in order to do this, 350 of the nodes will have handled the query, she said. "You're basically broadcasting your query to the whole network and this is a problem because as the network grows, the query traffic grows in proportion."

The researchers' algorithm, in contrast, takes almost twice as long, averaging eight steps to find 50 percent of the nodes in the same network, but because only one node handles the query at each step it is 40 times more efficient in terms of using network bandwidth, said Adamic.

The bandwidth efficiency of the researchers' algorithm addresses the key problem with these types of peer-to-peer networks: growing beyond 1,000 nodes.

This is currently a limit because the exponential bandwidth growth needed for broadcast querying causes the network to break up when the bandwidth requirements out run the network's slowest connections. "The network gets fragmented because the slow, or low bandwidth nodes such as the 56K modems can only hold something like 20 queries a second," said Adamic.

The researchers' algorithm scales well in power-law networks, requiring 12 steps to find 50 percent of the nodes in a 10,000-node network where nodes were aware of their neighbors' connections, said Adamic. When the algorithm did not target large hubs, but randomly chose a neighbor, the algorithm took an average of 40 steps for 1,000 nodes and 100 steps for 10,000 nodes, she said.

In contrast, in a non-power-law network whose nodes have roughly the same number of connections, the number of steps the algorithm takes scales exponentially, increasing from 100 to 1,000 as the number of nodes increases from 1,000 to 10,000.

The research is interesting, and potentially useful in the wide context of search engines on the Internet, said Alessandro Vespignani, a research scientist at the Abdus Salam International Centre for Theoretical Physics in Italy. "The paper shows for the first time that searching algorithms should be changed in order to take into account the complex connectivity nature of [power-law] networks," he said.

The researchers are currently working on using the algorithm in the field. "I think the first application of the search engine will be to help Gnuttella overcome its bandwidth barrier. Peer-to-peer networks are rapidly gaining in popularity, and the search methods are still being developed," Adamic said.

Adamic's research colleagues were Amit R. Puniyani of Stanford University, and Rajan M. Lukose and Bernardo A. Huberman of Hewlett-Packard's Sand Hill Labs. The research was funded by Stanford University and Xerox PARC.

Timeline: Now
Funding: Corporate, University
TRN Categories: Internet; Information Retrieval
Story Type: News
Related Elements: Technical paper, "Search in Power-Law Networks," posted on CORR at xxx.lanl.gov/abs/cs.NI/0103016

TRN

Content

Webs within Web Boost Searches

By Kimberly Patch, Technology Research News
November 13/20, 2002

Internet search engines regularly use information about the text contained in pages and the links between pages to return relevant search results because the approach works reasonably well, but less is known about why these relationships exist.

A researcher from the University of Iowa has expanded the utility of using text and links in search engines with a mathematical model that divides a large network like the Internet into small local Webs.

A Web crawler designed to completely traverse a small Web will provide more comprehensive coverage of a topic than typical search engines, according to Filippo Menczer, an assistant professor of management sciences at the University of Iowa. "My result shows that it is possible to design efficient Web crawling algorithms — crawlers that can quickly locate any related page among the billions of unrelated pages in the Web," he said.

Menczer's earlier work showed how similarities in pages' text related to the Web's link structure.

His latest work has expanded the concept by looking at a large number of pairs of pages from the entire Web and studying the relationships between three measures of similarity — text, links and meaning — across those pages. "A better understanding of the relationships between the cues available to us — such as words and links — about the

meaning of Web pages is essential in designing better ranking and crawling algorithms, which determine how well a search engine works," Menczer said.

The brute force approach gave Menczer enough data to uncover power-law relationships between textual content and Web page popularity and between semantic, or categorical, distance and Web page popularity. "From a sample of 150,000 pages taken from all top-level categories in the Open Directory, I considered every possible pair of pages, resulting in almost 4 billion pairs," said Menczer. The pattern would have been difficult to notice with smaller or nonrandom samples, he said.

Menczer used the data in a mathematical model that predicts Web growth, and showed that the model accurately predicted the way links are distributed in the Internet. "The Web growth model based on local content predicts the link... distribution," he said.

The model is based on the idea that Web page authors link to the most popular or important pages in their subject areas, said Menczer. The question is how they do this practically without a global knowledge of page popularity. Many existing models simply assume that a Web page author has knowledge of every Web site.

Menczer's model uses local content as a way to determine the probable distribution of links in a network. "In this sense the new model is more realistic because it is based on behavior that matches our intuition of what authors do," he said.

The model is relatively simple, Menczer said. "When you look at a new page, you link it to related pages which you know about with probability proportional to their... popularity," he said. The probability of linking between given pages decreases as the text similarities between them decreases, he said. The relationship between the probability of a link between pages and their text similarity follows a power-law, or exponential decrease.

The model, based on local knowledge, sees the Web as clusters of smaller webs of sites with similar topics. This bodes well for search engine developers, who can design Web crawlers to use textual and categorical cues to completely traverse a small Web in order to provide comprehensive coverage on a certain topic, according to Menczer.

The research should allow for ranking and crawling algorithms and more scalable search engines "where most pages of interest to a community of users can be located, indexed, and the semantic needs of users can be mapped into algorithms to destill the most related pages," Menczer said.

Menczer' research group is designing and evaluating topical Web crawlers, Menczer said. In addition, "we have some ideas on how to induce natural collaborative activities in communities of users that can emerge spontaneously in peer networks," he said. "Such activities will provide crawlers and indexers with rich contexts to improve their performance," he added.

Some progress in crawling and ranking is possible within a few years, but a full understanding of the complex inter-relationships between all sorts of information available on the Web will take longer to map out, he said.

Menczer is working on visual maps that will allow for a better interpretation of the relationships between text, links and the meaning of Web pages.

The work is useful and novel, said Shlomo Havlin, a physics professor at Bar-Ilan University in Israel. "It extends previous work on networks to [quantify] correlations between neighboring nodes. Such correlations have been found in realistic social and computer networks," he said.

The research adds to network models information that could improve researchers' understanding of aspects of networks like stability and immunization against software viruses, Havlin said. "This work extends the general body of research to include realistic features," he said.

Menczer published the research in the October 7, 2002 issue of *Proceedings of the National Academy of Sciences*. The research was funded by the National Science Foundation (NSF).

Timeline: > 3 years
Funding: Government
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Growing and Navigating the Small World Web by Local Content," *Proceedings of the National Academy of Sciences*, October 7, 2002

TRN

Web Pages Cluster by Content Type

By Kimberly Patch, Technology Research News
January 16, 2002

What makes the Web so useful is the vast amount of information it spans. And this is also what makes it so frustrating.

The challenge is indexing the Web in a way that allows people to find information quickly and painlessly. Scientists struggling with this problem have found that the Internet harbors more correlations among the types of information it holds than was at first apparent.

Information retrieval methods have long counted on correlations between word matches and meaning to find pages that are similar to each other.

A University of Iowa researcher has confirmed that there are also correlations between link distance and content, and link distance and meaning. "If two pages are separated by [only] a few links, then they are also similar in content and in meaning," said Filippo Menczer, an assistant professor of management sciences at the University of Iowa.

Untangling the correlations that exist among different aspects of the Web could be one key to better organizing its vast reaches.

The idea is that there are many notions of distance on the Web, and studying the relationships among these types of distance will provide cues to the relationships among Web pages, said Menczer. "It's like using cues in a physical environment. Suppose you are at a picnic in a park and you have to find the apple pie with your eyes closed. When the smell get stronger you know you're getting closer. So the strength of the smell signal is correlated with a physical distance," he said.

To verify the link-content correlation he measured the similarity of the words of many pairs of pages and the number of links that must be clicked to get from one to another. He also measured the link distances between pages that human experts had determined were similar in meaning.

"My results show that links... tell us a lot about the content and meaning of pages. This helps [us] understand why algorithms like Google's PageRank... work so well. They use links to estimate the meaning of pages," he said.

This strength of the correlations between links, text and meaning was surprising, said Menczer. "I found that beyond four or five links away, the probability [of finding] a relevant page is reduced to random chance," he said.

Menczer also found that the results varied depending on the type of domain he was measuring. "If you are browsing through Web sites of educational institutions, the signals are significantly more reliable than if you are surfing commercial sites," meaning the probability of finding a relevant page drops faster when you click away from commercial sites, he said. "In other words, you can get lost in cyberspace much faster when you're shopping online than when you are browsing a class syllabus," he said.

Taken together with two other recent findings in Web structure, the results could help build Web crawlers that do a better job of indexing, and cover more of the Web.

The Web is a small-world network, meaning it has a regular topology of pages clustered together, but also enough random links that they act as tunnels to reduce the average number of links between pages. This is the reason for the six degrees of separation phenomenon, which is that any person in the United States, or any Web page, can be reached from any other by making no more than six successive connections among people who know people, or among pages that are linked.

At the same time, it has become clear that finding these short paths to information is sometimes very difficult. The new correlations may help.

"The research I'm doing might shed light on this problem and help us understand whether it is theoretically possible to build efficient Web crawlers — agents that can find target pages in a reasonable time through local lexical and link cues," said Menczer.

Measuring and documenting the relationships between the structure of the Web and its content is clearly important, said Soumen Chakrabarti, an assistant professor of computer science at the Indian Institute of Technology in Bombay. "It has also been measured before, but not as systematically as in Menczer's paper," he said.

"Menczer takes an important step of modeling the coupling formally" and his model treats the link content relation more deeply than past research efforts, Chakrabarti added.

Menczer is working on Web crawlers that will take advantage of these topological findings. "The crawlers that now build a search engine's index... do not use knowledge about what the users are interested in," he said. Menczer's prototype Web crawler, dubbed MySpiders, is designed to better harness the clues in links and to integrate it with information from Web page content, he said.

This type of search engine could technically be ready for practical use within one or two years, said Menczer. The research was funded by the University of Iowa.

Timeline: 1-2 years
Funding: University
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Links Tell Us about Lexical and Semantic Web Content," posted on the arXiv physics archive at xxx.lanl.gov/abs/cs.IR/0108004. MySpiders Web crawler site: myspiders.biz.uiowa.edu

TRN

Ties That Bind Boost Searches

By Kimberly Patch, Technology Research News
March 13, 2002

Search for Michael Jordan on the Web and chances are you'll have a hard time finding the scientist of that name.

If you happen to know specifics about Jordan's branch of science, you can augment the search with a term like "Bayesian networks" in order to divert the inevitable flood of references to the popular athlete. But if you don't already know something about the scientist, you'll have to slog through many results to find the right guy.

Scientists from NEC Corporation's Research Institute are attempting to solve this basic Web search problem with an algorithm that takes advantage of the linking structure of the Web. "The solution... is to find an efficient and effective means of partitioning the Web in a manner that is independent of text content," said Gary William Flake, a research scientist at NEC Research.

The algorithm could lead to more precise search engines, according to Flake.

Organizing the Web in a way that is not dependent on the language contained within web pages is important because

language is ambiguous. "If we tried to cluster occurrences of 'Michael Jordan' by text content, we could run into many difficulties related to the ambiguity of languages," said Flake. In addition, "there may be many more 'Michael Jordans' out there, so we can't just arbitrarily choose to separate athletes from scientists," he said.

One way to organize the Web is to introduce some type of centralized hierarchy by hand, said Flake. Portals like Yahoo, for instance, employ people to index portions of the Web. But using people to comb through Web pages runs into a limit "related to how many humans you can get to do the dirty work," he said. "I suspect that there will never be a portal that indexes more than 10 million pages simply because it is too labor-intensive to do so."

Between automatic search engines and portals that use humans to organize content there is an implicit trade-off between reach and precision, said Flake. "On a search engine, you get back many results, but a lot of them are garbage. On portals, you get a small set of high-quality results, but with a lot of stuff missing," he said.

The algorithm attempts to span both approaches by providing a way to automatically organize the information based on the structure imposed on the Web by the way connections grow among pages.

Although it may seem random, the Internet actually contains a tremendous amount of structure, said Flake. The Internet is scale-free, meaning it has a few nodes, or pages that have many links to other nodes, and many nodes with just a few links. It is also a small-world network, meaning it has enough links, or shortcuts between large nodes, or hubs, that it is possible to get from one node to any other by taking just a few hops. This phenomenon is also responsible for the six degrees of separation found in social networks.

The algorithm connects the Web's linking structure to communities like science, finance, health, education, or recreation. A Web community has more links among community members than it does outside of the community. There are several algorithms that track links to find sets of related pages, but they're either inefficient or cover only a portion of the Web, according to Flake.

The key to coming up with an efficient algorithm for identifying communities was to start with a set of hand-picked seed sites, he said. "The entries within a portal category typically make excellent seeds."

The algorithm looks to the Web's structure to calculate a community that contains the seed sites, said Flake. "In this way, we can improve the [reach] of a portal while preserving the precision," he said.

The algorithm then does a popularity-like ranking within the community.

The group's continuing work crosses three distinct areas, said Flake. "On the mathematical front, we [are] refining our algorithm to make it more efficient and more accurate." The group is also working to mathematically characterize the properties of communities, he said.

"On the engineering front, my group is working on improving our hardware and software infrastructure so that our algorithms will scale up to the case where we're dealing with billions of Web pages," he said.

On the scientific front, the researchers are studying how large-scale communities on the Web relate to one another and how new communities emerge, said Flake.

"For the Web, I'm hoping that we can turn this into niche search engines, automatic portals, smart content filters, and even user agents," said Flake.

The community algorithm may eventually also be applicable to other types of networks, said Flake. "I would like to see if we can effectively apply this to other data sets that have a network-like structure, such as those found in biology," he said.

The researchers have found a novel way to infer content from links without using textual information, said Filippo Menczer, an assistant professor at the University of Iowa. The work "pushes the envelope with respect to just how much useful information we can extract from the self-organizational Web hypertext — the fact that people selectively link their Web pages to other, probably related pages," he said.

The method could prove useful for augmenting search engines, said Menczer. "For example, Google burned the competition when it first introduced link analysis into the realm of commercial search engines. There are still many clues hidden in the Web that can be exploited, both in text and link information," he said. "It's an incredibly rich environment and we have only begun its exploration."

One issue with this particular method is how much time the algorithm takes. Because the Web is so large, it may be impractical to crawl the Web to the depth it requires, Menczer said. "An interesting direction for this work is to explore efficient ways to apply the idea. While the algorithm does not make any use of text information, it is possible that using text to guide the preliminary crawls," may speed community identification, he said.

The researchers are aiming to make a viable content filter that they can demonstrate within a year, said Flake. "Later, on the order of two years from now, we may introduce a niche search engine that ultimately customizes itself to individual users," he said.

Flake's research colleagues were Steve Lawrence, C. Lee Giles and Frans M. Coetzee. They published the research in the March, 2002 issue of the journal *IEEE Computer*. The research was funded by NEC.

Timeline: 1 year, 2 years
Funding: Corporate
TRN Categories: Internet
Story Type: News

Related Elements: Technical paper, "Self-Organization of the Web and Identification of Communities," *IEEE Computer*, March, 2002

TRN

Net Scan Finds Like-Minded Users

By Kimberly Patch, Technology Research News
May 7/14, 2003

When you search for information on the Web, chances are you aren't alone — there are like-minded groups of users across the Web searching for the same sorts of things.

Researchers from the University of Chicago have shown that it is possible to identify these groups by analyzing browsing patterns, even in networks as far-flung as the Web.

The researchers' method of graphing information across data distribution systems like the Internet shows that, given a large enough sample, computer users can be grouped according to their common interests based only on their requests for data. "One of the first questions we asked was is the group-based collaboration of scientists mirrored somehow in their usage of data," said Adriana Iamnitchi, a researcher at the University of Chicago.

The answer turned out to be yes, across all types of group-based interests. "Communities as heterogeneous as the Web seem to show this pattern of having users naturally group in interest-based groups," she said.

The information-request graphing method can be used to design scalable, adaptive methods for locating and delivering data, said Iamnitchi. The method could theoretically be used by anyone, including ecommerce vendors, to target communities of interest.

The researchers are working on using the patterns to design more efficient services for resource-sharing environments like Grid computing, Iamnitchi said.

Grid software coordinates a few or even hundreds of computers across networks like the Internet to piece together compute power and resources like databases into powerful virtual computers; the combined resources can speed up scientific and engineering applications like time-consuming equations and three-dimensional simulations.

The researchers found the data-sharing relationship pattern while looking for a way to leverage characteristics of the Grid computing community to make that type of computing more efficient, according to Iamnitchi. "Our idea was to... design mechanisms [that are] able to cope efficiently with large and dynamic numbers of resources — data files, computers, and storage space for results," she said.

One typical characteristic of the community that uses Grid computing is they tend to collaborate, said Iamnitchi. When the researchers analyzed traces of scientific computations from a high-energy physics collaboration that spanned 18 countries and involved 70-odd institutions and thousands of physicists, they found that the patterns of collaboration were mirrored in scientists' data requests.

The researchers looked at the relationships that formed among users based on the data they were interested in. "We captured and quantified these relationships by modeling the system as a data-sharing ... graph whose nodes are the data consumers in that system," said Iamnitchi. Nodes, or people, who requested a given number of the same files within a given time were connected.

In an analysis of six months worth of scientists' requests for data, the researchers found that group-based collaboration is visible in the way information is requested, said Iamnitchi. "Scientists form groups of interest based on the data they used," she said. The researchers found the same pattern in a larger analysis of general Web requests.

The pattern of similar requests shared the small-world characteristic common in many networks, including the way data is arranged in networks like the Internet.

In small-world networks, it is possible to get from one node to any other node by traversing relatively few links. Social networks, with people as nodes and relationships as links, and the Web, with pages as nodes, and links between pages as links, are also small-world networks.

Looking at small-world topologies is not a novel idea, but the method of extracting a graph from an arbitrary data-sharing relationship and using it to study these structures is, said Filippo Menczer, an assistant professor of management services at the University of Iowa.

Data request patterns have been analyzed previously, but in different ways — to examine the popularity distribution of Web requests or to study the most efficient way to cache Internet traffic. In contrast, the Chicago researchers' analysis uncovered relationships between users based on their common interests in data.

The method is potentially useful, especially because a graph can be made from any Web usage log, said Menczer. "Any Webmaster can do this."

The method may be useful for discovering clusters of users who have interest in a certain type of data, Menczer said. "Ecommerce vendors are currently using collaborative filtering techniques, which are related to this," to do so, he said. The method can also be used for distributed caching and broadcasting, similar to the services offered by Akamai Technologies Inc., he said.

The researchers are now making the method more efficient for resource-sharing environments like Grid computing, said Iamnitchi. "We are currently looking... to design mechanisms to locate resources," she said. "The ultimate goal is to provide scalable, adaptive mechanisms [that are] able to deal with variations in resource participation."

The resource location mechanisms could be ready to use within two years, Iamnitchi said.

Iamnitchi's research colleagues were Matei Ripeanu from the University of Chicago and Ian Foster from Argonne National laboratory. The research was funded by the National Science Foundation (NSF).

Timeline: 2 years
Funding: Government
TRN Categories: Internet; Distributed Computing
Story Type: News
Related Elements: Technical paper, "Data-ring Relationships in the Web," posted on the arXiv physics archive at arXiv.org/abs/cs.NI/0302016

TRN

Conceptual Links Trump Hyperlinks

By Kimberly Patch, Technology Research News
July 10/17, 2002

The art of navigating through physical space has improved considerably in the millennia since people learned to get their bearings from the stars, making getting from point A to point B a fairly straightforward experience. The nascent art of navigating through cyberspace, however, is often much more frustrating.

It's clear that people form some kind of mental model as they click through cyberspace. What's less clear is exactly how the mental model works, and how Web sites can be designed to fit it.

Researchers from Kansas State University have found that some common approaches to Web design contribute to Web users' frustration. The good news is that there's a fairly easy fix — more closely align the way Web pages link to each other with the way the concepts within Web pages relate to each other. Another study has found that the consistency of navigation options affects how people organize these concepts.

In line with all the old metaphors about the Internet being an information highway, much of today's Web site design assumes these mental models form a type of spatial map. "If this were true, we would expect people would be able to remember where things are located and how the Web site is organized," said J. Shawn Farris, a Kansas State University researcher.

The researchers set out to see how well people could mentally map Web sites by asking 40 subjects to explore one of four versions of a Web site, then getting the subjects to draw a map of how the sites' pages were linked together. The different versions contained the same information, but the page links were organized differently.

The results showed a surprising gap between the organization of the Web sites and the resulting mental models. "According to what is frequently assumed... we expected people to be able to remember how the Web site was organized," said Farris. "We expected them to form a survey-like cognitive map of the Web site. They did not," he said.

Instead of depicting the different ways the pages were linked together, the drawings showed how the information related conceptually, said Farris. In fact, the mental models of people who looked at the four different Web site matched. "All participants drew the same thing," said Farris.

The research showed that people tend not to remember how pages link together, but they do remember how the concepts embodied in the pages relate to one another, said Farris. People "remember how content is organized," he said.

Although the findings do not line up with the conventional wisdom of Web design, they do make sense given what is known about cognitive psychology, said Farris. "When we remember things, we tend to form conceptual associations," he said. "For example, if we have just finished talking about dinosaurs, we tend to more quickly recall the names of specific dinosaurs even though we did not talk about those specific dinosaurs in our conversation."

The researchers are looking into an area that needs to be explored, and the results make sense, according to David Danielson, a researcher at Stanford University. "That navigators conceptualize the Web spatially has long been a common assumption — there are 'routes' between Web pages, some pages are... 'landmarks', and site 'maps' allow the user to 'survey' the site. Even the widely-used term 'navigate' implies that point of view," he said.

The question is, to what extent does the site designer impose a particular metaphor on the user, said Danielson. "Site designers may have more power over a Web visitor's way of thinking than they are sometimes given credit, or blame, for," Danielson said.

The researchers' challenge to the pervasive, spatial way of thinking "will hopefully open up a more lively debate," said Danielson. The work "reinforces an idea that every information architecture should pay heed to: the site's connectivity should match the way its information is interrelated," he said.

According to Danielson's own work on Web design, the extent to which navigation options change as the user moves through the site also affects her mental model of the site. "A user can view a site as a congregation of loosely-connected facts, tightly related topics, or somewhere in between," he said.

Danielson tracked people as they used the Web to find information, and varied the navigation context of pages while keeping all other site attributes the same. He found that when people's navigation options changed dramatically as they moved through a site, they became more disoriented, and also viewed site topics as less tightly-related.

When the navigation option changes were more subtle, however, they saw connections between pages more easily, said Danielson. "Users [tended] not to notice the option

changes, and probably got a chance to see connections rather than differences," he said.

In the end, if Web developers can design pages that more closely align with mental models, people will be better able to find what they want on the Web. "Helping a user develop an accurate mental model of your site answers a few critical questions she'll have when faced with a link:

'Where will it take me? Where will I be able to go from there? Have I already gone down this path? How much of that have I seen?' It gives her predictive power," said Danielson. "When she doesn't have that power, you can see it in her behavior: she backtracks more often," he said.

Farris' research colleagues were Keith S. Jones and Peter D. Elgin. Their paper detailing the research is slated for an upcoming issue of the journal *Interacting with Computers*. The research was funded by the University.

Timeline: Now
Funding: University
TRN Categories: Computer Science; Human-Computer Interaction; Internet
Story Type: News
Related Elements: Technical paper, "Users' Schematic of Hypermedia: What Is so 'Spatial' about a Website?," *Interacting with Computers*, in press

TRN

English Could Snowball on Net

By Ted Smalley Bowen, Technology Research News
November 21, 2001

The Internet's ability to connect a wide range of cultures would seem to bode well for diversity of all sorts.

But, while the technology is relatively neutral, the influences of political and economic power have made the Internet a virtual English-language empire.

Researchers from the Tel Aviv University and the University of California at Berkeley have teamed up to gauge the nature of the relationship between linguistic patterns and Internet content.

Early returns from the work imply that English content will continue to dominate the Internet, although other studies predict different scenarios.

Currently about 70 percent of Internet content is in English, but only about 44 percent of Internet users are native English speakers. Worldwide, native Spanish speakers outnumber native English speakers, and the number of native Chinese speakers more than equals that of both groups. English dominates online because it was established early on as the lingua franca of the wired world.

The imbalance reflects a first-mover advantage that is common in networks of all kinds, according to Neil Gandal, an associate professor of economics at Tel Aviv University in Israel.

In this case, the language of Shakespeare, Mark Twain, H.L. Mencken, and Yogi Berra benefits from the snowballing effect of a popular medium attracting more users simply because it's popular. The language's popularity spurs more people to learn English, which increases incentives for content providers to cater to an English-speaking audience, which in turn makes it all the more popular.

The researchers examined whether these first-mover effects dictate that English will simply gain momentum and remain the primary online language, prompting even more people to learn it, or whether the demographic and economic realities of a polyglot world will turn the tide.

This question is especially pertinent because Internet use among non-native English speakers is growing at a faster rate than that of native English speakers. By 2003 only 29 percent of Web users will be native English speakers, according to one estimate.

The researchers analyzed the surfing habits of a usefully bilingual population — Canadians in the province of Québec. As of 1996, roughly 5.7 million Québec citizens counted French as their mother tongue, about 600,000 cited English, and about 60,000 listed both.

The researchers looked at users' overall time online and time spent at each of seven types of sites: retail, business and finance; entertainment, news, sports and technology; education; portals, searches and directories; services, including ISPs, careers, and hobbies; government; and adult.

To get a rough breakdown by language of the content surfed, the researchers wrote a spider program that identified the languages of the approximately 40,000 Quebecois URL domains visited.

The researchers compared the overall Internet use of the three linguistic camps by type of sites, regardless of the content language, and then looked at which factors determined the percent of the time devoted to English language sites.

The native English speakers visited English content sites 87 percent of the time and stayed online about 35 percent longer than their French-speaking neighbors. The native French speakers, however, surfed in English a still considerable 64 percent of the time.

The differences also narrowed with age: younger native French speakers looked at more English content than their elders.

The finding that native French speakers are hurdling the linguistic barrier and turning to English sites for content not available in French is evidence that English's first-mover advantage is still snowballing, according to Gandal. These network effects are likely to continue to favor creating content in English and to lower incentives to do so in French, he said.

These preliminary results also indicate that the Internet is increasing the incentive for non-native English speakers to

learn English as a second language, which could in turn promote English as a global language, according to Gandal.

In addition, although automatic translation technologies may eventually break down linguistic barriers, they are currently too limited to be a likely influence on the choice of content language, said Gandal. "Translation is very difficult because of the subtlety involved in the use of language," he said.

Computer-generated translation does work well for finding simple information like a train or airline schedule or the location of a particular office, but does not convey more complicated communications like disease diagnosis or an explanation of how to make a retail purchase, said Gandal. "We don't think that they will play a prominent role in the choice of language content in the foreseeable future."

The issue of language representation on the Internet is a contentious one, and is complicated by widespread financial stakes and cultural implications. The researchers' conclusions contradict those of the Foundation for Networks and Development, a private regional development organization in the Dominican Republic.

The current predominance of English on the Internet is largely due to the network's American origins and because the first wave of users worldwide is more likely to speak English as a second language, said Daniel Pimienta, director of the Foundation.

The foundation's statistics show that this is changing, he said. For instance, three years ago 75 percent of Web pages were in English, but that number has dropped to 50 percent today. In addition, the number of English Web pages as a percentage of the population of the world that speaks English as a native or second language is falling relative to Spanish, French, Italian and Portuguese, he said.

As the Internet's population becomes more diverse and an increasing percentage of its users lack English skills, the early predominance of English will continue to fade, he said. "As the Internet evolves toward a more balanced geographical [distribution] and a more balanced socio-economic distribution, the dominance of English will more and more appear as a transitional phenomenon and the representation of language in the Net will tend to become closer to the natural representation of the language in the world."

As this happens, however, English will retain a special role in bridging communities whose native languages are different, he added. "This is and will remain the case of English, but also of Spanish, French, Arabic and Chinese."

Under this scenario, monolingual native English speakers may be more likely to pick up another tongue, Pimienta said. "The Internet will probably represent a strong asset for the language training industry to add a second language to native English speakers."

The Tel Aviv and Berkeley team's choice of a mostly bilingual population like Quebec's makes it harder to gauge the factors driving the choice of language on the Internet, Pimienta said. That population is able to navigate in English, while 90% of the world population does not understand English, he said.

The Tel Aviv and Berkeley researchers are currently working on a model designed to distinguish among cultural and economic factors driving the spread of English and those effects specific to the Internet, Gandal said.

One goal is finding how closely the use of English online will hew to the demographic and economic realities of English speakers. "The question is whether the percent of Internet content in English will reflect... or... greatly exceed the percentage of native English speakers around the world, weighted by purchasing power," said Gandal.

The researchers plan to delve into data for all of Canada in an effort to quantify factors like the number of Internet pages read or transactions conducted that would justify continued use of and investment in a particular language, Gandal said. "The model will need to distinguish between adults who find it harder to learn a new language... and children who find it easier," and therefore get more out of the experience, he said.

The researchers' updated model will also help quantify the strong network effects favoring development in English and drawing the best bells and whistles to English sites which, at least initially, place non-English sites at a disadvantage.

As more precise language identification software emerges, the researchers will be better able to determine the breakdown of pages visited according to content language, according to Gandal.

Gandal's research associate was Carl Shapiro of the University of California at Berkeley. They presented the work last month at the Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy in Alexandria, Virginia. The research was funded by the UC Berkeley.

Timeline: Now
Funding: University
TRN Categories: Internet, linguistics
Story Type: News
Related Elements: Technical paper, "The Effect of Native Language on Internet Usage", Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy, October 27-29, 2001, Alexandria, Virginia

TRN

Vulnerabilities

Hubs Key to Net Viruses

By Kimberly Patch, Technology Research News
November 7, 2001

When the Internet linked distant computers 30 years ago, its founders were probably not thinking about protecting the machines from infecting each other. Today's exponentially larger Internet, however, is vulnerable to software viruses in much the same way that large, crowded human populations are more likely to fall prey to biological viruses.

The Internet has a scale-free structure, meaning it has a few pages, or nodes with many connections to other pages and many with just a few connections. Researchers looking into how bits of disruptive code spread on the Internet have found that this structure isn't conducive to the conventional practices of inoculating large populations.

The researchers did, however, find an inoculation strategy that promises to protect computers more effectively.

When they applied an immunization strategy that's commonly used for biological populations to a simulated scale-free network, it simply didn't work, said Alessandro Vespignani, a research scientist at the Abdus Salaam International Center for Theoretical Physics in Italy.

The researchers inoculated progressively larger numbers of nodes, expecting the epidemic to eventually die out, he said. It did not even when they inoculated more than 90 percent of the nodes, he said. "Surprisingly, in scale-free networks we observed that infection survived... in the presence of massive vaccination campaigns involving the majority of the population. We realized that random... schemes were practically useless in scale-free networks."

The Internet is generally more vulnerable than human populations because the connections among computers are both more numerous and structured differently than many of the human connections that allow viruses to spread. Scale-free networks have some nodes — large portals, for instance — that contain more connections to other pages than even the most widely-traveled people could possibly have with other people.

The researchers eventually caused the epidemic to die out by targeting nodes that had a high number of connections rather than inoculating individuals randomly.

Using this scheme, the researchers sharply lowered the network's vulnerability to epidemic attacks, Vespignani said. "We have tested this recipe on a real map of the Internet [with] a targeted immunization involving all the most-connected individuals. In this case, by immunizing [less] than one percent of the total population, the cyber infection cannot propagate," he said.

The research explains why, though antivirus software is very successful in protecting individual computers, it does not prevent computer infection from becoming endemic. "The 'I love you' virus is still in the top list of most frequent viruses more than a year after its introduction... because the global implementation of antivirus [software] is practically equivalent to a random... vaccination," Vespignani said.

Ironically, this scheme could also be useful in the biological world where some of the paths viruses take to propagate in a human population have some similarities to the Internet. A map of human sexual relations, for instance, has scale-free properties, said Vespignani. The research implies that epidemics spread this way could be prevented more effectively by targeted vaccination of the few promiscuous individuals, he said.

This type of targeted vaccination would also prove to be much cheaper than the random kind, Vespignani said. "Instead of massive vaccination campaigns, we can think of identifying the network connectivity hierarchy." Controlling the hubs that spread the infection more quickly is both more effective and requires relatively few inoculations, he said. "The strategy is... particularly convenient in terms of economical and practical resources."

The problem in both the Internet and biological networks that harbor a scale-free nature is identifying the large hubs, said Vespignani. "The difficulty... is... detailed knowledge of the network connectivity. This is not always possible for privacy and economical reasons. It is very difficult to obtain a complete map of the Internet because many providers do not want to share publicly their information. As well in the case of sexual diseases we have to rely on people's concerns about their own sexual habits," he said.

This strategy "looks reasonable. It is consistent with my experience," said Gene Spafford, a computer science professor at Purdue University. "I'm surprised no one else has noted this property in research... either in networks or in epidemiology," he said.

One complication that the model leaves out is the notion of workgroups, or local area networks where each machine is connected to all the other machines in that group, and an infection of one infects all the others, Spafford added.

It is hard to estimate when the research could be used to actually inoculate networks, said Vespignani. "The use of these results is strictly related to social factors — individuals' privacy — and the existence of control agencies." These make estimating the time frame difficult, he said.

Vespignani's research colleague was Romualdo Pastor-Satorras of the Technical University of Catalonia in Spain. The research was funded by the European Community, the Spanish Ministry of Education and Culture, the Abdus Salaam International Center for Theoretical Physics (ICTP) and the Technical University of Catalonia (UPC).

Timeline: Unknown
Funding: Government, Private
TRN Categories: Internet
Story Type: News

Related Elements: Technical paper, "Optimal Immunization of Complex Networks," posted in the Los Alamos physics archive at arXiv.org/abs/cond-mat/0107066

TRN

Five Percent of Nodes Keep Net Together

By Kimberly Patch, Technology Research News
May 23, 2001

Because the Internet is a distributed network with no central server directing information flow, there are many potential paths from any given point on the network to any other point. This makes it a robust network that is difficult to shut down.

The Internet is also a scale-free, or power-law network, meaning it harbors a small number of very large hubs with many connections to other nodes, and a large number of nodes with only a few connections. This concentration of connections, a trait the Internet shares with large social and biological networks, makes it more vulnerable to intentional attack, however, than a network with more evenly distributed node sizes.

Researchers from Bar-Ilan University in Israel and Clarkson University are examining just how vulnerable the Internet's scale-free nature makes it. Knowing more about scale-free networks' vulnerabilities may point the way to both protecting the Internet from attacks and providing better strategies for attacking biological networks in order to fight disease.

While the Internet is made up of computers that are connected via communications lines to other computers, a typical biological scale-free network is made up of the molecules a cell uses. In this case, the network connections are interactions among molecules. The large hubs in a cell's chemical communications network include water and the cellular fuel ATP, which are used in many more reactions then most of the molecules it uses.

The researchers work shows that large scale-free networks are fairly impervious to random node breakdowns, but if large hubs are targeted methodically, even large scale-free networks can be broken up into separate islands. "We've studied the problem mathematically. According to our findings, while networks like the Internet are resilient to random breakdown of nodes, they're very sensitive to intentional attack on the highest connectivity nodes," said Shlomo Havlin, a physics professor at Bar-Ilan University.

This is because a scale-free network's stability depends on the state of its large hubs, he said.

In scale-free networks as large as the Internet, "there are just enough high connectivity nodes to keep the network connected under any number of randomly broken nodes," he said. "A random breakdown of nodes will leave some... highly connected sites intact, and they will keep a large portion of the network connected," he said.

An attack that targets about five percent of these highly connected sites, however, has the capacity to totally collapse the Internet, "very rapidly [breaking] down the entire network to small, unconnected islands," containing no more than 100 computers each, Havlin said.

The researchers cannot pinpoint the breakdown threshold any more precisely than near five percent, Havlin noted, because the exact distribution of nodes on the Internet can only be roughly estimated.

To find the threshold, the researchers used a branch of mathematics known as percolation theory, which was originally developed to predict how much oil can be pumped from a reservoir. "Since oil can only flow through holes in the ground, this is similar to data flowing through... computers on the Internet," said Havlin.

Another way to picture percolation theory is to draw a square lattice of dots on a piece of paper. If you remove a small number of the dots, you can still connect the rest of the dots around the ones you have removed. "However, after removing the critical fraction [of dots] there's no continuous paths from side to side," said Havlin.

In terms of the Internet, "as long as we're above the threshold, there will be a large connected structure with size proportional to that of the entire Internet. Below the threshold, there will only be small unconnected islands of sizes in the dozens [of nodes] each," he said.

The researchers' work offers the theoretical basis for calculating the threshold for the breakdown of any complicated network, said Albert-László Barabási, a physics professor at the University of Notre Dame. "By offering a method to calculate... the number of nodes required to be removed in order to destroy the network by breaking it into isolated clusters, it will be of great use [in] fields ranging from Internet research to drug delivery, where the goal is, [for example,] to destroy some microbes by gene removal. I expect this result will have a lasting impact on our understanding of the resilience of complex networks in general," he said.

The researchers' aim is to find ways to design networks that are more resilient to both random error and intentional breakdown, said Havlin. The work may also lead to better understanding of network traffic and virus propagation on the Internet, he said.

Havlin's research colleagues were Reuven Cohen and Keren Erez of Bar-Ilan University in Israel, and Daniel ben-Avraham of Clarkson University. They published the research in the April 16, 2001 issue of *Physical Review Letters*. The work was funded by the Bar-Ilan University and the Minerva Center.

Timeline: Now
Funding: Institutional, University

TRN Categories: Networking
Story Type: News
Related Elements: Technical paper, "Breakdown of the Internet under Intentional Attack," *Physical Review Letters*, April 16, 2001

TRN

Net Inherently Virus Prone

By Kimberly Patch, Technology Research News
March 21, 2001

The Internet's sheer size and large central hubs make it an efficient communications network, but those same traits make it vulnerable to the uninvited bits of code that are the computer's equivalent of biological viruses.

Two physicists have applied their understanding of condensed matter physics, which examines the complex, collective behavior of matter, to mapping how viruses traverse the Internet's complicated labyrinth of connections.

What they have found is that the Internet's efficient communications structure may make it vulnerable to even the weakest of viruses.

Standard epidemiological models look at how virulent biological viruses are. The more virulent, or easily spread a virus is, the larger the risk that it can spark an epidemic. If the virulence falls below a certain threshold, however, the infection will die out exponentially fast and therefore cannot spread fast enough to become a threat.

In order to study virus spread within the Internet, the physicists took into account the network's scale-free structure. The Internet harbors a few extremely large hubs, or nodes with huge numbers of connections and, many nodes with only a few connections.

In contrast, the hubs in social connections are more limited in size. "Real viruses can be transmitted only by close physical contact, and so diffuse in the community in a series of short hops between infected and uninfected individuals," said Alessandro Vespignani, a research scientist at the Abdus Salam International Centre for Theoretical Physics in Italy. "The crucial difference is that computer viruses spread on the Internet, which has a very special branching structure so on the Internet viruses can always pervade the system," he said.

The researchers found that this type of structure allows the epidemic threshold to fall below zero, meaning that no matter how low a virus' virulence, it won't necessarily die out. "Strikingly, we found that the Internet lacks any epidemic threshold. The Internet is prone to the spreading and the persistence of infections [no matter how low their] virulence," said Vespignani.

The Internet's structure also explains why computer viruses affect only portions of the Internet. Despite spreading easily and being able to survive for a long time, any single virus has a low probability of infecting the bulk of the Internet, according to the researchers.

Ironically, "the ideal world for data sharing and fast communications... is also an ideal environment for viruses, which easily find... ways to rapidly infect new hosts through the intricate digital highways," he said. "The connections between computers on the Internet have enormous fluctuations and intricate structure that has to be included in the theoretical and experimental study of digital epidemics."

The lack of a threshold is surprising and potentially important for other scale-free networks as well, said Albert-László Barabási, a physics professor at Notre Dame University.

Network models have until now shown that viruses invariably die out if they're not too contagious, said Barabási. "However, [these models were] based on... outdated ideas on the topology of real networks," he said. In recent years it has become increasingly clear that many networks, including the Internet, are scale-free and have inhomogeneous topologies dominated by a few highly connected hubs, he said.

The lack of a threshold for virus spreading in scale-free networks "is highly unexpected and it will have a significant effect on a number of fields," said Barabási.

The model could be used to understand epidemic dynamics in scale-free networks like "food webs, power grids and social networks," said Vespignani. It could also be applied to problems like the spread of polluting agents, he said.

There's another issue worth looking at that may make understanding how computer viruses spread more complicated, however, said Jon Kleinberg, an assistant professor of computer science at Cornell University. "It's a subtle issue. What is really the network on which these viruses spread?" he said.

Instead of the full, scale-free network of the Internet, many computer viruses actually spread on subsets that look more like social networks that have limitations on how large a hub can be, he said. It is very common, for instance, for a computer virus to spread by sending itself to the e-mail addresses listed in an infected machine's address book.

While the large, central hubs on computer networks can have tens of thousands or even millions of connections, the size of even the largest e-mail address books are limited. "The real network on which viruses spread is an invisible network of who talks to whom sitting on top of the Internet, and that's a network that we have less ability to measure at the moment," said Klineberg.

The researchers are fine-tuning their models to see how effective things like immunization could be on the Internet. "We're introducing more details and realism in the model. For instance, we're considering the presence of immunization, latency times, [and] detailed Internet maps," Vespignani said.

The researchers are headed toward a general theory of epidemiology and complex networks, he added. Such a model

could help in devising algorithms to protect the Internet from a virus epidemic, he added.

Ultimately, the Internet needs a global immunization organization in order to establish the optimal policies of immunization and antivirus implementation, said Vespignani. "We claim that the Internet needs a digital immune system... that automatically detects and submits viruses to some central control laboratory," he said.

Vespignani's research colleague was Romualdo Pastor-Satorras of the Polytechnic University of Catalonia in Spain. The research was funded by The International Centre for Theoretical Physics, the Polytechnic University of Catalonia, the European Community Network and the Spanish Ministry of Education and Culture.

Timeline: Unavailable
Funding: Government, Private
TRN Categories: Internet; Networking; Cryptography and Security
Story Type: News
Related Elements: Technical paper, "Epidemic Spreading in Scale-Free Networks," scheduled to appear in the April 2, 2001 issue of *Physical Review Letters* and posted on the Computing Research Repository (CORE): arXiv.org/abs/cond-mat/?0010317

TRN

Network Comparisons

Social Networks Sturdier Than Net

By Kimberly Patch, Technology Research News
February 12/19, 2003

Although many types of networks, including biological networks, social networks, and the Internet, have a lot in common, when you get right down to who is connecting to whom, social networks follow different rules.

A researcher from the Santa Fe Institute has found that social networks are assortative, meaning people who are social gravitate toward others who are social. This is very different from many other types of networks.

The nodes of social networks are people. People who already have connections like to associate with other nodes who have connections, said Mark Newman, now an assistant professor of physics at the University of Michigan.

In contrast, non-social networks like the Internet, World Wide Web, and biological networks are disassortative, meaning highly-connected nodes tend to connect to nodes that have few connections, said Newman.

There's a "big difference between social networks and all other kinds of networks," said Newman. This was somewhat unexpected, and it has several ramifications, he said.

In social networks, where popular people are friends with other popular people, diseases spread easily, said Newman. At the same time, however, this type of network has a small central set of people that the disease can actually reach. "They support epidemics easily, but... the epidemic is limited in who it can reach," he said.

The opposite is true for the Internet, the Web and biological networks, said Newman. This makes these types of networks more vulnerable to attack than social networks are.

The implications for vaccinating people and for protecting networks like the Internet against attacks are not good, according to Newman. The networks that we might want to break up, like social networks that spread disease, are resilient against attacks; but the networks that we wish to protect, like the Internet, are vulnerable to attack, said Newman.

Social networks hold together even when some of the most connected nodes are removed. This may be because these nodes tend to be clustered together in a core group so that there's a lot of redundancy, according to Newman. This means that vaccination and similar strategies are less effective than in other types of networks.

Attacks on the largest nodes of disassortative networks, however, affect the network as a whole more because the connections are more broadly distributed across the network. "This suggests that if nodes were to fail on the Internet, it would have a bigger effect on the performance of the Net than we might otherwise expect," he said. "In a way, it is telling us that the Internet is fragile."

Newman found that the number of highly-connected nodes that need to be removed to destroy disassortative networks is smaller by a factor of five or 10 than the number needed to destroy assortative networks.

The technical challenge to doing the research was the computer simulations, said Newman. They "involve some tricks that I had to work out specially for this study," he said. The mathematics involved and the computer simulations all tied together to give a clear picture of what is going on, he said.

The model could eventually be used to better understand how diseases spread, said Newman. "[It could] suggest better strategies for preventing their spread," he said.

The model could also be of use in the Internet, he said. "Ultimately the aim... is to understand how network systems work, and how the structure of the network affects their performance, for example, how the structure of a social network affects the way societies work," he said.

The model could be used in epidemiological work now, said Newman.

He published the research in the October 28, 2002 issue of *Physical Review Letters*. The research was funded by the National Science Foundation.

Timeline: Now, 3 years
Funding: Government
TRN Categories: Networking; Physics
Story Type: News

Related Elements: Technical paper, "Assortative Mixing in Networks," *Physical Review Letters*, October 28, 2002

TRN

Motifs Distinguish Networks

By Kimberly Patch, Technology Research News
November 27/December 4, 2002

There are many types of networks in the world — computer webs like the Internet, connections among components in electronics, relationships among friends and acquaintances, transportation grids, food relationships among animals, connections among neurons, and interactions among genes.

Scientists from the Weizmann Institute of Science in Israel and Spring Harbor Laboratory have shown that it is possible to categorize networks by looking at certain recurring circuits, or motifs, within the networks. "The motifs are small, local, wiring patterns that occur throughout the network," said Uri Alon, a senior scientist at the Weizmann Institute of Science.

Identifying and examining these motifs can help explain how networks function, Alon said. "The motifs allow us to break up the network into building blocks," he said. "This gives the hope that understanding the function of each motif would allow us to build up an understanding of the entire network behavior," he said.

Understanding network motifs could contribute to better Internet search engines, a better understanding of networks within cells, which could help in curing disease, and a better understanding of social networks, which could help heal societal rifts.

Each type of network appears to have its own characteristic set of motifs, said Alon. "This is probably because they are functional units important to whatever function the network was designed or evolved to perform," he said.

The feedforward loop, or filter motif, for example, is common in networks of neurons, but is relatively rare and therefore not a motif in food webs, said Alon. A feedforward loop consists of network nodes x, y and z, in which x has connections to y and z, and y also has a connection to z. In food webs, where nodes are animals, this pattern is carried out only by omnivores (x) that both eat another animal (y) and the food (z) that that animal eats, he said.

The researchers' work is complementary to a line of research that looks for broad patterns across networks. That line of research has shown, for example, that many networks, including the Internet, have scale-free and small-world attributes.

Scale-free networks contain a few nodes that link to many other nodes, and many nodes with few links. Small-world networks contain short paths between large nodes, allowing the networks to be traversed using fewer hops between nodes; this broad pattern appears across social networks and the Internet and is responsible for the well-known six-degrees-of-separation effect.

To find the network motifs, the Weizmann researchers compiled databases of networks, including all known transcription interactions in bacteria, and wrote algorithms that analyzed information. "We had to make up efficient algorithms that [could] handle networks with millions of nodes and count their subgraphs," or repeating connection patterns, he said.

Transcription occurs whenever a cell needs to make a protein. Transcription molecules, which are also proteins, direct the copying of a portion of DNA that provides a physical blueprint for the needed protein.

The researchers' approach showed that groups of networks share certain motifs. "For example, seven different food webs share the same two motifs," Alon said. Those motifs are a chain, where one type of prey eats another, which eats another, and a diamond-shaped pattern, where one type of prey eats two others, which both eat a fourth type of prey.

These motifs are very different from those that occur in biochemical networks, which share the feedforward loop and a bi-fan pattern, where x and y can each transmit information one-way to z or w, according to Alon.

And "these are again different [from] the motifs found in the World Wide Web," Alon said. The researchers found five motifs in the Web: a fully-connected triangular relationship that shows two-way connections among x, y, and z; a modified feedback loop where there are two-way connections between x and y, and y and z, but a one-way relationship from z to x; and three increasingly less-connected relationships.

A modified feedback loop appears on the Web when there is a two-way link between a university homepage (x) and a department page (y), a two-way link between a department page (y) and a lab page (z), but, because the university does not have room to list all the labs in its homepage, only a one-way link from z to x.

The presence of many two-way connections in the Web may reflect a design weighted toward providing short paths between related pages, according to Alon.

The researchers also found that networks that perform information-processing share similar motifs, even if those networks are otherwise quite different, according to Alon. "Sometimes two networks from completely different fields, made up of completely different elements, show the same motifs," said Alon. "This occurs, for example in transcription networks and neuron networks, even though one describes proteins and genes within a yeast cell, and the other describes neuron wiring in a worm," he said.

The similarity in motifs probably reflects a fundamental similarity in the design constraints of the two types of networks, according to Alon. Both types of networks carry information from sensory components to components that carry out a task. In a transcription network, transcription proteins regulated by biochemical signals communicate with

genes that build proteins; in a neural network motor neurons transmit signals to muscles.

One possible function of this type of motif is to activate output only if the input signal is persistent, and to allow a rapid deactivation when the input goes off, according to Alon.

The researchers also found that human engineering uses design rules similar to those used by evolution, said Alon. "Both converge again and again on a small set of useful circuits" or motifs, he said. "Electronic chips are built of recurring circuits such as operational amplifiers and filters. Engineers love to use these elementary circuits because they are robust and... plug-and-play. Evolution did not go to engineering school, and yet still uses similar design principles," he said.

The network motifs should provide strong clues about what makes a complex network tick, said Alon. For example, the feedforward loop and amplifier motifs of biological networks will be very useful for designing nanoscale devices that need to compute, he said. "The circuits favored by biology will be the ones that work [well in] engineering on the nanoscale," he said.

In computer science, understanding the basic recurring structure of the Web may contribute to better ways to search and design networks.

And in medicine, the hardest diseases to cure have to do with networks — the forces governing when a cell divides and when it dies, for instance, said Alon. "Doctors attempting to fix the cell are working today without a blueprint of its networks," he said.

More information about networks could even help heal societal rifts, he said. "Working in Israel, one hopes that understanding the basic elements of social networks may one day work to heal the cycles of violence that occur in societies and nations," he said.

The researchers work is interesting because "it tries to make connections between the different types of complex networks beyond the power-law results that we have seen thus far," said Filippo Menczer, an assistant professor of management services at the University of Iowa. "It goes beyond global link analysis such as the studies which unveiled the scale-free/power degree distribution of many complex networks including the Web, and starts focusing on more local structures," he said

The work suggests that the very small patterns of connectivity that appear much more or much less frequently than you'd expect in random networks must serve some purpose, said Menczer. "The patterns found for the Web [, however,] are not too surprising and may [simply] reflect common behaviors of authors' linking to related pages," he said.

If the method can be applied to larger motifs, it could shed light on unknown functional mechanisms of networks, Menczer added.

The method can be applied to any network now, Alon said.

The researchers' are working on extending the method to analyze cell-wide biochemical networks, said Alon. "We would like to obtain the concepts and tools needed to build a blueprint of a cell," he said.

They are also working on figuring out why certain types of motifs appear in some networks but not others. "Can we divide the world of networks into universal classes, each optimized to perform different tasks?" said Alon. "The far-off goal is to reach a unifying theory of evolved and designed" networks, he said.

Alon's research colleagues were Ron Milo, Nadav, Kashtan, Shalev Itzkovitz and Shai Shen-Or at the Weizmann Institute of Science, and Dmitry Chklovskiie at Cold Spring Harbor Laboratory. They published the research in the October 25, 2002 issue of the journal *Science*. The research was funded by the Israel Science Foundation, The Human Frontiers Science Foundation and the Minerva Foundation.

Timeline: Now
Funding: Government, Private
TRN Categories: Physics; Internet
Story Type: News
Related Elements: Technical paper, "Network Motifs: Simple Building Blocks of Complex Networks," Science, October 25, 2002

TRN

Network Similarities Run Deep

By Kimberly Patch, Technology Research News
January 3, 2001

Networks exist everywhere — in living systems and social groups as well as among computers. What's more, there are similarities among networks as diverse as the World Wide Web and the chemical goings on of a bacterium.

A pair of scientists working to map out what networks have in common have come up with a theory that uses the tools of calculus to uncover the ways that networks develop. The work could lead to both better search algorithms for the Web and better means of disrupting the chemical networks of cells for medicinal purposes.

The work builds on previous research that found that the links in networks are not random. Whether the network is of chemicals in a cell, interactions among people, or Web pages, once the network grows to a certain size, it develops hubs where links congregate. These larger hubs exist in all networks the researchers have examined, said Albert-László Barabási, a physics professor at the University of Notre Dame.

"The World Wide Web, the actor network, the science citation network and many other networks in nature follow the same pattern... show some kind of self organizing process ... that is different from simple random networks," said Barabási.

In the actor network, each actor is a node, and actors are linked to each other when they play together in a movie. The similar science citation network maps out scientists who work together. This corresponds to Web pages connected by links and chemicals connected by reactions in natural systems.

In the actor and scientists networks, certain people emerge who are linked to many more people than the average. "If [an actor] is very popular, they will acquire links at an even higher rate because every casting director would like to have them in the movie. Some of these processes can contribute to make the hubs even larger," said Barabási.

The Web has larger hubs as well. In living systems, certain chemicals, like water and ATP, play a role in large numbers of reactions, while a very large number of chemicals have just one or two roles. A simple bacterium, for example, uses between 1000 and 2000 chemicals, Barabási said.

This pattern is a continuum, with a few extremely large hubs stepping down rapidly into a very long tail of nodes with an increasingly smaller number of links each. This pattern corresponds to the continuum model, which allows the researchers to "mathematically describe how these networks are growing... and to predict the characteristics of the network," said Barabási.

The Notre Dame researcher's work is one of two ways to describe systems, said John Klineberg, assistant professor of computer science at Cornell. A discrete model looks at nodes and links as distinct objects. A continuous model looks at them as approximations. Each model has its pluses and minuses.

"Fundamentally the network is a discrete object — it has an actual number of nodes, so it's not really infinitely divisible, but ... once you approximate things ... as continuous you have at your disposal the tools of differential equations — calculus. And so the question is going to be is your system big enough that you can really model it continuously," said Klineberg.

If the models were describing how water boils rather how networks are constructed, a discrete model would have the impossible task of tracking the trajectory of each molecule as the water heated up, while the continuum theory would find aggregate properties of the system in order to derive statistically that the water will boil at a certain temperature after a certain amount of heat is added, said Klineberg.

The continuous approach is well-suited to describe a pot of boiling water because it contains about 10 million trillion molecules. With that many objects "you don't lose much by approximating," said Klineberg.

At this point the Web is too big to model discretely, but still much smaller than a pot of boiling water. "We're at a point where it's interesting to look at both approaches... and it's reassuring that the two approaches are getting similar answers," said Klineberg.

The Notre Dame researchers have used the Continuum Theory to look at several traits of networks to see if the way networks develop affects this basic model. "Now that we understand that [networks] are vaguely similar to each other, they follow the same mathematical principles, [we're trying] to understand the details of this process," Barabási said.

They have found that the basic structure of networks makes them fairly impervious to random errors, but if several of the largest hubs are simultaneously attacked, the whole network is at risk. "If you go for the big nodes, not randomly, but you take simultaneously down, say, one percent of the biggest nodes in the system, then the system will break apart into pieces," said Barabási.

Networks have several other traits in common, he said. Networks are often growing systems. "You're always adding new Web pages, new people to society," he said. In addition, when a new link is added, the most connected nodes are generally preferred.

In a biological system, however "we don't know why these parallels appear because they're very tricky ... there is evolution which is supposed to optimize the system [but] it's not clear" how such a complicated process affects the network, Barabási added.

It is also apparent that networks share a small diameter, said Klineberg. "You can get from one point to any other point with very few hops — that's also coming out as we look at this."

There are also ways networks differ, Klineberg said. For instance, while a social network is symmetrical, meaning relationship links are usually two-way, Web links are asymmetrical. "Lots of people link to Yahoo's homepage... but Yahoo doesn't reciprocate," he said.

The Notre Dame researcher's work is already starting to be applied Web searching, Barabasi said. "The Google search engine is using some of these principles already. The more we understand about the topology of the system, the better we can design tools to work with the systems, whether it's a search engine or it's a [chemical] that's trying to kill a cell," he said.

Barabási's research colleague is Reka Albert from the University of Notre Dame. They published a technical paper about the research in the December 11, 2000 issue of *Physical Review Letters*. The work was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Networking; Internet
Story Type: News
Related Elements: Technical paper, "Topology of Evolving Networks: Local Events and Universality," *Physical Review Letters*, December 11, 2000; Technical papers posted at www.nd.edu/~network/papers.htm; Related technical paper, "Network Robustness and Fragility: Percolation on Random Graphs," *Physical Review Letters* December 18, 2000

TRN

Index

TRN's Making The Future Report is published 10 times a year by Technology Research News, LLC. Each 20- to 40-page package assesses the state of research in a field like biochips, data storage or human-computer interaction.

Single reports are $300 to $500. A one-year subscription is $1,600. To buy a report or yearly subscription, go to www.trnmag.com/email.html.

We welcome comments of any type at feedback@trnmag.com. For questions about subscriptions, email mtfsubs@trnmag.com or call (617) 325-4940.

Technology Research News is an independent publisher and news service dedicated to covering technology research developments in university, government and corporate laboratories.

Kimberly Patch
Editor
kpatch@trnmag.com

Eric Smalley
Editor
esmalley@trnmag.com

Ted Smalley Bowen
Contributing Editor
tbowen@trnmag.com

Chhavi Sachdev
Contributing Writer
csachdev@trnmag.com