THE ELEVENTH LINK

The Awakening Internet

WHEN PAUL BARAN REGISTERED a week late for his first computer science class at University of Pennsylvania, he knew that he had already-missed the first lecture, but he was not too worried. Not much is done in the first class anyway. So he showed up for the second class, on Boolean algebra, the mathematics behind computer logic. As he recalls, "The instructor went up to the blackboard and wrote '1 + 1 = 0.' I looked around the room waiting for someone to correct his atrocious arithmetic. No one did. So I figured that I may be missing something here, and I didn't go back." Yet, he did revisit the subject ten years later, on his fourth job after graduation. This time he faced a different problem: He was way too early.

Barely thirty and only a few months into his new job at RAND Corporation, Baran was given the prodigious task of developing a communication system that would survive a nuclear attack. In 1959 the possibility of a Soviet nuclear warhead's falling from the sky was not mere science fiction but an appropriately feared potential war scenario. Baran's employer, a California think tank founded in 1946 to provide the intellectual know-how for the military's nuclear buildup, had considerable expertise in developing war scenarios and potential disaster outcomes. Such grim tasks as foreseeing and detailing the death of millions from a nuclear attack were never a source of good press, often tarring the company with Dr. Strangelove's brush. Baran's assignment, to

develop a survivable communicator system, was par for the course at RAND. Baran took his job seriously, and in a twelve-volume series of RAND Memorandums he meticulously described the vulnerabilities of the existing communication infrastructure and proposed a better one—the Internet.

Baran saw the vulnerability of the command system of the 1950s hidden in the topology of the existing communication network. Since a nuclear strike handicaps all equipment within the range of detonation, he wanted to design a system whose users outside of this range would not lose contact with one another. Inspecting the communication systems of that time, he saw three types of networks (see Figure 11.1). Baran discarded the starlike topology, concluding that "the centralized network is obviously vulnerable as destruction of a single central node destroys communication between the end stations." Baran saw the current system as a "hierarchical structure of a set of stars connected in the form of a larger star," offering an early description of a «cale-free network. With incredible insight, he found this topology too centralized to be viable under attack. In Baran's mind the ideal survivable architecture was a distributed meshlike network, similar to a highway system, redundant enough so that even if some nodes went down, alternative paths maintained the connection between the rest of the nodes.

An enduring myth alleges that the Internet was designed to survive a Soviet nuclear strike. It is true that Baran's main motivation was to design a system that could not be taken out by the Soviet nuclear arsenal. But in the long run his ideas and innovations were all but ignored by the military. As a result the topology of today's Internet has little to do with his vision. Yet the topological change advocated by Baran was not the reason everyone from the military to industry vehemently opposed his design. The objection was to his proposal to break the messages into small packets of uniform size capable of traveling independently of one another along the network. This could not be achieved with the existing analog communication system. Thus he advocated a switch to a digital system. This step was too difficult for AT&T, the communication monopoly of his time, to absorb. Therefore, AT&T's Jack Osterman quashed Baran's vision when he declared, "First, it can't

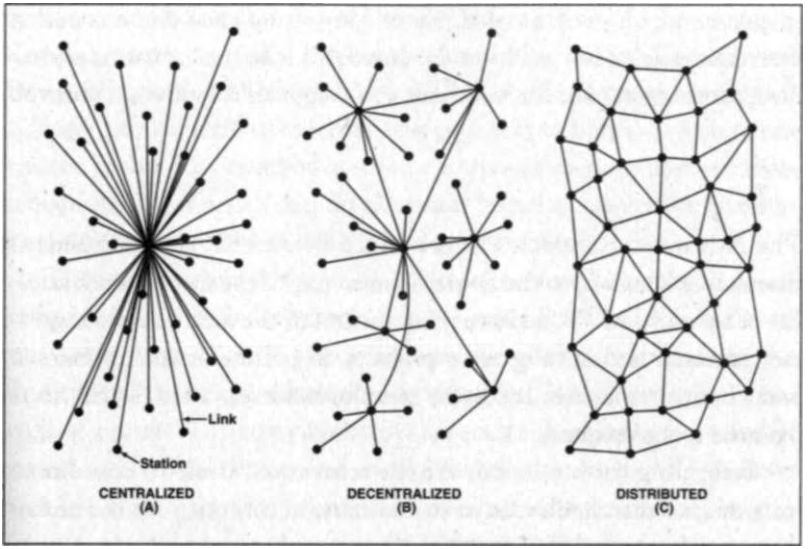

Figure 11.1 **Paul Baran's Networks.** *In 1964, Paul Baran began thinking about the optimal structure of the Internet. He suggested that there were three possible architectures for such a network—centralized, decentralized, and distributed—and warned that both the centralized and decentralized structures that dominated communications systems of the time were too vulnerable to attack- Instead, he proposed that the Internet should be designed to have a distributed, mesh-like architecture. (Reproduced with permission of Paul Baran.)*

possibly work, and if it did, damned if we are going to allow the creation of a competition to ourselves." Baran's ideas, defeated at every step by industry and the military, were rediscovered only years later, when the Advanced Research Projects Agency, not aware of his results, independently constructed the same vision. By that time, however, the Internet was well along its course of development.

Understanding the topology of the Internet is a prerequisite for designing tools and services that offer a fast and reliable communication infrastructure. Though human made, the Internet is not centrally designed. Structurally, the Internet is closer to an ecosystem than to a Swiss watch. Therefore, understanding the Internet is not only an engineering or a mathematical problem. In important ways, historical forces

shaped its topology. A tangled tale of converging ideas and competing motivations left their mark on the Internet's structure, creating a jumbled information mass for historians and computer scientists to unravel.

1.

The Advanced Research Projects Agency, or ARPA, was President Eisenhower's answer to the Soviets' launching of the first Sputnik satellite. Originally ARPA had sweeping control of the most advanced military research and development projects, in particular the antimissile and satellite programs. It lost its muscle, however, after NASA took over the space program.

Struggling for a mission, ARPA reinvented itself to coordinate long-range research relevant to the military, in contrast with the immediate developmental projects that different military agencies were handling themselves. The Internet entered the picture around 1965 or 1966, when Bob Taylor, the director of ARPA's computing program, suddenly became concerned with a huge waste of federal resources he had just discovered.

In the 1960s, ARPA was already funding computer research in a big way. This indeed required considerable investment—with the PC revolution decades away, computers cost anywhere from half a million to several million dollars. ARPA had several such monsters in its research portfolio, hosted by research labs around the country. The problem was that even computers in the same room could not talk to each other. Tapping into the computing power stored at other ARPA-supported sites was out of the question. Bob Taylor had a brilliant idea: To stop this waste, why not link these incompatible machines somehow? In February 1966, after presenting his vision to Charlie Herzfeld, ARPA's director, he walked away with a fresh million in his budget and a new sense of mission.

The idea of connecting computers also occurred to Donald Davies, director of computer science at Britain's National Physical Laboratory[1] in Teddington, a town within commuting distance of London. Working hard to turn his idea into reality, Davies reinvented packets and packet

switching well before learning of Baran's preexisting work. His group presented these concepts at a 1967 symposium in Gatlinburg, Texas, introducing his and Baran's ideas to the ARPA-supported research group. It suddenly became clear to everyone that packet switching over faster lines was the technology required to create a truly efficient communication network. Finally Baran's decade-old vision began to materialize. And so the network that today we call the Internet was born.

The word *Internet* is often used to describe everything related to our online universe, including computers, routers, optical cables, and even the World Wide Web. Here we will use the word to refer only to the physical infrastructure connecting computers. The Internet is a network of routers that communicate with each other through protocols envisioned by Paul Baran and made possible thanks to ARPA's deep pockets. Ironically, the principles directing today's Internet match Baran's original vision in every respect except the guiding principle that motivated his work: undercutting vulnerability to attacks. Baran's distributed highwaylike network could have become a reality only if the Internet had continued to be regulated and maintained by the military. The Internet, however, took on a life of its own.

2.

In the computer science community Bill "Ches" Cheswick, a researcher at a Lucent/Bell Labs' spin-off called Lumeta, is best known for his work on firewalls and computer security. But the public increasingly recognizes him for the colorful Internet maps he and Hal Burch, also at Lumeta, produce and sell through Peacockmaps.com. The millennium map, depicting the Internet's topology on January 1, 2000, shows a **dense,** entangled forest of routers and links, a network of considerable beauty. Its complexity is matched perhaps only by the human brain. There is an important difference between the two, however. Whereas **the** human brain's size has been stagnating for centuries, the Internet continues to grow exponentially, without any sign of slowing down.

Cheswick is far from being a lone scientist with artistic aspirations. He **is in** illustrious company. DARPA, the successor to ARPA, is currently

spending millions of dollars on research groups around the United States to do just what Cheswick is doing: map the Internet. The most visible of these projects is the Cooperative Association for Internet Data Analysis, or CAIDA, an Internet tomography collaboration hosted by the University of California at San Diego, whose main goal is to monitor just about every characteristic of the Internet from traffic to topology. Across the Atlantic but only a click away, Martin Dodge, a researcher at the Center for Advanced Spatial Analysis at University College London, hosts Cybermaps.com, a colorful Website collecting a stunning body of maps visualizing the Internet.

Would it ever occur to you to meticulously draw a map of your watch, the Pentium chip in your computer, or the car youvdrive everyday to work? Hardly. If you really want to know what is under the hood, you could contact the manufacturer for the car's blueprint. Engineers prepare hundreds of maps before building each watch, chip, or car, detailing not only every component, but the location of and the relationship between each piece, as well. But today, when the Internet is the workhorse of the American economy, we still do not have a detailed map of it. Since the National Science Foundation relinquished its stewardship of the Internet in early 1995, no central authority has controlled or documented its growth and design.

Today the Internet evolves based on local, distributed decisions on an "as needed" basis. Everyone, from corporations to educational institutions, adds nodes and links without needing permission from a central authority. There is no single network either. Independent but interlinked networks coexist and operate, going by such names as WNET, vBNS, or Abilene.

You would think there was someone out there who, if necessary, could shut the whole thing down. Wrong. While you could persuade an institution to close down the portion of the network under its authority, no single company or person controls more than a negligible fraction of the whole Internet. The underlying network has become so distributed, decentralized, and locally guarded that even such an ordinary task as getting a central map of it has become virtually impossible.

3.

There are important practical reasons for seeking a global Internet map. Without knowing the Internet's topology it is impossible to design better tools and services. The current Internet protocols were developed with a small netwotk and 1970s technologies and needs in mind. As the network grew and new applications emetged these protocols have often fallen short of our desires. Indeed, most of today's use of the Internet was unimaginable by those who designed the basic infrastructure, which is still in place. For example, e-mail was born when an adventurous hacker, Rag Tomlinson, working at BBN, a small consulting firm in Cambridge, Massachusetts, figured out how to modify the file transfer protocols to carry mail messages. For a long time Tomlinson kept quiet about his bteakthtough. When he first showed it to one of his colleagues, he warned him, "Don't tell anyone! This isn't what we're supposed to be working on." E-mail leaked out, howevet, and became one of the dominant applications of the early Internet.

The same is true of the Wotld Wide Web. The infrastructure was never prepared for it. It is an excellent example of a "success disaster" the design of a new function that escapes into the real world and multiplies at an unseen rate befote the design is fully in place. Today the Internet is used almost exclusively fot accessing the World Wide Web and e-mail. Had its original creators foreseen this, they would have designed a very different infrastructure, resulting in a much smoother experience. Instead we find ourselves locked into a technology that adapts only with great difficulty to the booming diversity and demand imposed by the increasingly creative use of the Internet.

Until the mid-nineties all research concentrated on designing new protocols and components. Lately, however, an increasing number of researchers are asking an unexpected question: What exactly did we create? While entirely of human design, the Internet now lives a life of its own. It has all the characteristics of a complex evolving system, making it mote similar to a cell than to a computer chip. Many diverse components, developed separately, contribute to the functioning of a

system that is far mote than the sum of its parts. Therefore, Internet researchers are increasingly morphing from designers into explorers. They are like biologists or ecologists who ate faced with an incredibly complex system that, fot all practical purposes, exists independently of them. The mystery is a bit deeper than that, however. While biologists have spent decades figuring out what proteins look like and how they intetact with each othet, all details tegatding the Internet's components are fully available to the Internet tomographer. What neithet computer scientists nor biologists know is how the large-scale structute emerges once we put the pieces together.

4 .

Vern Paxon and Sally Floyd, computet scientists at the International Computer Science Institute Centet fot Internet Research in Berkeley, California, in an influential and much quoted 1997 paper, identified our limited knowledge of the netwotk topology as the main obstacle toward a better understanding of the Internet as a whole. Two years later thtee Greek computer scientist brothers, Michalis Faloutsos of the University of California-Riverside, Petros Faloutsos of the University of Toronto, and Chtistos Faloutsos of Carnegie Mellon University, made a surprising discovery. They found that the connectivity distribution of the Internet routers follows a power law. In their seminar paper "On Power-Law Relationship of the Internet Topology" they showed that the Internet, a collection of routers linked by various physical lines, is a scale-free network. Their discovery had a simple message that quickly penetrated the research community: All tools used to model the structure of the Internet before 1999, based on ideas rooted in random networks, were simply wrong.

The Faloutsos brothers were unaware of the parallel discoveries of power laws in the World Wide Web topology. Combined with these developments their finding acquited a new meaning, removing the Internet from the world of random netwotks and dropping it into the colorful zoo of scale-free topologies. This was rathet unexpected. Aftet all,

the Internet is comprised of physical lines and routers. It is all hardware. How could these costly and heavy copper and optical connections follow the same rules as humans do when establishing their weightless social links or adding URLs to their Webpage?

5.

In October 1969 Charley Kline was asked to arrange the first computer-to-computer message through an ordinary telephone line. Working as a programmer in the UCLA lab of Leonard Kleinrock, he was part of a project attempting to connect to the only other existing Internet node located at Stanford University. After establishing the connection, Kline started by typing "login." He typed *l* and got the echo from Stanford confirming that the letter had been received. He proceeded with o and again received the appropriate echo. Then he ventured to g. However, that was too much for the young system to absorb, and the computer crashed, killing the connection as well.

The connection was quickly reestablished, and after the UCLA and Stanford nodes were firmly in place many others joined in. According to John Naughton, author of A *Brief History of the Future,* the University of California-Santa Barbara and the University of Utah got the third and the fourth nodes in November and December 1969, respectively. The fifth was delivered to BBN, a Massachusetts consulting firm, early in 1970, together with the first cross-country circuit—a second line connecting the machines in Los Angeles to BBN's in Boston. By the summer of 1970, nodes six, seven, eight, and nine had been installed at MIT, RAND, System Development Corporation, and Harvard. By the end of 1971 the Internet consisted of fifteen nodes; by the end of 1972 it had thirty-seven. As Naughton puts it, "The system was beginning to spread its wings—or, if you were of a suspicious turn of mind, its tentacles."

As you may have noticed, the Internet follows the classical scenario °**f** a growing network. Today, two decades later, it continues to expand node by node—the first and necessary condition for the emergence of a

scale-free topology. Preferential attachment, the second condition, is more subtle, however Why would anyone link his or her computer to any routet other than the nearest one? After all, laying down a longer cable is more expensive.

It turns out that the length of cable is not the limiting factor determining the growth or stagnation of the Internet. When an institution decides to link its computers to the Internet, it has only one parameter in mind: cost of communication. Regatding bandwidth, the measure of how many bits a connection can carry each second, the closest node is often not the best choice. Going a few extra miles could provide access to fastet routers.

Routers offering more bandwidth likely have more links as well. Thus, while shopping for a good place to link, netwotk engineers inevitably gravitate toward the mote heavily connected access points. This simple effect is a possible source of preferential attachment. We do not know for sure whethet it is the only one, but preferential attachment is unquestionably present on the Internet. This was first demonsttated by Soon-Hyung Yook and Hawoong Jeong, both wotking in my research group, when they compared Internet maps recorded at several months' time intervals. Charting how the Internet grows node by node they found quantitative evidence that nodes rich in links acquire more links than nodes with a few links only.

Growth and preferential attachment should be sufficient to explain the scale-free topology discovered by the Faloutsos brothers. On the Internet things are a bit more complicated, however While not the primary consideration, distance does matter Undeniably, it is more expensive to lay down two miles of optical cable than half a mile. We must also take into considetation that nodes do not appear randomly across the map. Routers are added where there is a demand for them, and demand depends on the number of people wanting to use the Intetnet. Thus there is a strong correlation between population density and the density of the Internet nodes. The distribution of routers on the map of North America forms a fractal set, a self-similar mathematical object discovered in the 1970s by Benoit Mandelbrot. Therefore, when trying to model the Internet, we must simultaneously acknowledge the

interplay of growth, preferential attachment, distance dependence, and an underlying fractal structure.

Each of these forces alone, if taken to the extreme, could destroy the scale-free topology. For example, if the length of the wire were the main consideration when deciding where to link, the resulting network would have an exponential degree distribution, developing a topology very similar to the highway system. But the amazing thing is that these coexisting mechanisms delicately balance each other, maintaining a scale-free Internet. This very balance of power is the Internet's own Achilles' heel.

6.

MAI Network Services, a small Internet service provider headquartered in McLean, Virginia, owns several high speed Internet routers linked to the giant networks owned by Sprint and UUNet. On the morning of Friday, April 25, 1997, MAI released a routing table update for its routers. Routers shepherd packets they receive toward their destination by matching the address on the packet with a routing table. These routing tables are the roadmaps of the Internet. As the network topology is constantly changing, the routing tables are also periodically updated. At 8:30 **A.M.** MAI broadcast the updated routing information to its own routers. Because of an incorrect configuration, the update did not stop at the borders of MAI but escaped and rewrote the routing tables of a large number of routers owned by Sprint and UUNet. It instructed them to send all traffic to several MAI routers.

It was like watching water burst from a broken dam, destroying everything in its path. MAI watched in horror as all Internet traffic was suddenly redirected towards it. Because it never had the capacity to handle even a fraction of this flood, MAI turned into a black hole, absorbing packages at an incredible rate. Forty-five minutes later the company was forced to shut itself down to stop the damage. In the meantime Internet providers helplessly watched all their traffic get sucked into the black hole created by the faulty reconfiguration. Sprint recovered only after it manually changed all the routing tables it

owned, as did many of the big and small Internet providers affected by the problem.

Thanks to the quick resolution and the relative youth of the Internet, the world paid little attention to the event. However, it offered a vivid demonstration of the speed by which errors propagate on the Net: Within minutes of its release the misconfigured routing table was part of several large networks, triggering a classic example of a cascading failure.

Paul Baran had a very specific threat in mind when he designed the prototype of the Internet. He anticipated Soviet nuclear warheads hitting intelligence and military headquarters, potentially leading to complete information and communication loss. Neither he nor_the early Internet pioneers considered the possibility that one day people from any country in the world could have access to the infrastructure. For many years the United States resisted sharing the technology with countries deemed nonfriendly. I experienced that myself, as the much hated COCOM list officially excluded Hungary from the Internet until the fall of the Berlin Wall. The Internet was too contagious to be halted by such artificial barriers, however. Thanks to the ingenuity of local system managers, many eastern European universities had been regularly communicating via e-mail with their Western colleagues well before the restrictions were lifted. Today virtually every country on Earth is connected to the Internet. This open access policy brought along unexpected dangers and vulnerabilities as well, increasingly threatening our interlinked world.

One of the United States' busiest nodes, owned by AT&T, is a highly guarded subterranean facility in Schaumburg, Illinois, a Chicago suburb. This and several similarly well protected key nodes offer a false sense of security that the Internet cannot be broken by intentional attacks. The increasingly understood interplay between the network architecture and the protocols presents a different picture, however. A few well-trained crackers could destroy the net in thirty minutes from anywhere in the world. There are many ways to accomplish this, from breaking into the computers running several key routers to launching denial-of-service attacks against the busiest nodes.

The Code Red worm, which spread like a virus and infected hundreds of thousands of computers worldwide in the summer of 2001, is a good example of a technology that could achieve just that destruction. At first it appeared to be a harmless virus, since it did not damage its host. But after sitting dormant for days, it suddenly turned all infected computers into zombies, simultaneously throwing traffic at whitehouse.gov. Code Red was only a proof-of-principle demonstration of what automated viruses could achieve. More sophisticated versions could result in unparalleled damage. Disabling a few major nodes would not be sufficient to break the network into pieces, but the cascading failure of other routers resulting from the redirection of traffic to smaller nodes would finish the job.

Most crackers or hackers with the know-how would have no interest in taking the whole Internet out. A successful attack would take away their favorite toy, denying them access to the Net, as well. So a large-scale action taking on the entire Internet would never be the work of true hackers. But it could easily be the goal of rogue nations and terrorists. Understanding the Internet's topology will help us protect it.

7.

On August 30, 2001, National Public Radio aired a five-minute segment about our latest research, published the same day by the British journal *Nature.* It was not the first time that our work had been featured in the media. But the next morning, staring in disbelief at the project's Website counter, which had registered over 10,000 hits overnight, I realized things were a bit different this time. My e-mailbox was crowded with uncountable messages. Most were positive. Some, however, were rather scary. "Stay the hell out of my computer!" wrote a senior officer in a company developing deterrence programs. "I'd hate to see another Eastern European CompSci person tossed in jail by the US Federal Government," concluded another less than friendly note, reminding me of the recent arrest of a Russian hacker by U.S. authorities. "[I] request that you assure us that no computers on our networks have been, or are currently being, targeted by this program," wrote the CEO of a company in

Norway. "I remind you that any unauthorized use of resources located at these IP addresses is illegal and may result in legal action and demands for compensation." How could a research paper intended fot an academic audience and published by one of the most prestigious scientific journals create such a fierce and immediate reaction?

James McAdams, head of the Department of Government and International Studies at Notre Dame, had a great idea in early 2000. He assembled seven professors from all different departments, including economics, physics, law, chemical engineering, computer science, and Asian languages, to discuss in an informal setting the impact of the Internet on everything from democracy to teaching. Meeting once a month for lunch or breakfast, we took turns suggesting discussion topics and assigning reading materials, covering issues from cyberlaw to social movements' on the Web. During one such breakfast meeting computer scientist Jay Brockman mentioned that the Web is a computer, metaphorically speaking. His comment left me puzzled. To be sure, the Internet is comptised of computers that can exchange Webpages and e-mail messages. But this limited, user-driven communication does not yet make the Wotld Wide Web a single computer.

Could we do something to change this? Could we make computets drive each other's activity? To get started, could I force any computer out there to do computation on my behalf? Now this was an interesting question that I was willing to entertain. We ended up forming a tiny research group to try to address it. Btockman and I were soon joined by Vincent Freeh, an expert on Internet protocols, and my longtime collaborator Hawoong Jeong. After many discussions and tutorials on how computers communicate, a simple but controversial idea emetged: *parasitic computing.*

Sending a message through the Internet is a sophisticated process regulated by layers of complex protocols. Fot example, when you click on a URL to view a Webpage, your request is broken into small packets that are then carried to the computer owning the Webpage. There the request is reconstructed and interpreted, prompting the distant computer to send you the requested Web document. Therefore, such a seemingly simple task as clicking on a URL involves a significant

amount of computation along the way. Parasitic computing exploits this setup by forcing computers to perform computation at the command of a master host by merely engaging the computers in communication. To achieve this we disguised complex computational problems as legitimate Internet requests. When a computer received a packet, it performed a routine check to ensure that the packet had not been corrupted during its journey. While doing the math, it solved a problem of interest to us, encoded into the packet.

Our implementation of parasitic computing demonstrated that we can enslave computers located thousands of miles away, forcing them to perform computation on our behalf. This fundamental vulnerability of the Internet raised a barrage of computational, ethical, and legal questions. What if someone improves the method, making it efficient, and starts using it on a grand scale? Who owns the resources that are made available to anyone through the Internet? Could this mark the birth of the Internet computer? Will there be a new intelligent being at the end of this road?

Taken to an extreme, parasitic computing suggests that in the future computers could swap information and services on an as-needed basis. Right now communication within a chip is orders of magnitude faster than communication across the Internet. With broadband communication on its way, the gap will shrink. Soon it will start making perfect sense to ask other computers to chip in their unused resources to solve complex problems that cannot be addressed by a single computer or research group. On a smaller scale this possibility has already been exploited by SETl@home, a Berkeley-based project that harbors the unused time of millions of PCs to search for extraterrestrial intelligence.

The SETI model requires your voluntary collaboration. Most of us are just simply too lazy to go along. If, however, protocols allowing service and information swapping become the norm, vast unused resources could be tapped. Along the way the Internet might become independent of human supervision, since it can shepherd most of the information and resources it needs to solve specific problems. This could have unforeseen impact on the Internet's topology as well, giving self-

organization an even bigget role. I can imagine a time when, aftet getting an answer to a question from your Web browser, neither you nor your computer will know for sure where it came from. Aftet all, do you know where the letter A is stored in your brain?

8.

Our skin is a unique piece of engineering. It has the ability to measure and sense changes in temperatute and movement of air; it can size up objects and identify their make. It achieves all of this with the help of a huge number of tiny integrated chemical sensors that talk to each other through the nervous system. As Neil Gross pointed out jn *Business-Week,* a skin of similat sensitivity is enfolding the earth right now. Millions of measuring devices, including cameras, microphones, thermostats and tempetature gauges, light and traffic sensors, and pollution detectors, are popping up everywhere, feeding information into increasingly fast and sophisticated computets. Experts ptedict that by 2010 thete will be around 10,000 telemetric devices for each human on the planet. This number is not particulatly significant in and of itself—we've had sensots for a long time, ranging from surveillance cameras in supermarkets to cat detectors buried in the pavement at traffic signals that switch the lights at the intersection. What is changing is that fot the first time these various sensors are feeding information into a single integrated system. There will soon be over 3 billion Internet-connected cell phones and close to 16 billion Internet-connected computers embedded in everything from toasters to fashion designs. The tiny sensors of this planetary skin will spy on everything from the environment to out highways and bodies. Most importantly, however, they are all connected. Our planet is evolving into a single vast computer made of billions of interconnected processors and sensors. The question being asked by many is, when will this computet become self-aware? When will a thinking machine, orders of magnitude faster than a human brain, emetge spontaneously from billions of interconnected modules?

It is impossible to ptedict when the Internet will become self-aware, but clearly it already lives a life of its own. It grows and evolves

at an unparalleled rate while following the same laws that nature uses to spin its own webs. Indeed, it shows many similarities to real organisms. Just like the millions of reactions taking place in a cell, terabytes of information are passed along its links every day. The surprising thing is that some of this information is very difficult to find. That brings us to yet another network: the World Wide Web.

THE TWELFTH LINK

The Fragmented Web

SCIENCE FICTION WRITERS and visionaries, whose books I consumed as a child, made me believe that by the turn of the century human-looking robots would handle all mundane tasks. Yet we entered the new millennium without such humble servants having appeated on the scene. Or perhaps the robots have arrived quietly. They do not have the shining golden exterior of the always worried C-3PO, nor can they produce the joyful whistle of R2-D2. They wisely avoid sharing the crowded Euclidean space with us, whete real estate is at a premium. The robots of the twenty-first century are invisible and immaterial. They have taken up residence in the virtual world, which allows them to hop with enviable ease from continent to continent. Staring at your computer screen won't reveal these robots. But if you take the time to inspect carefully your computer's log files, which keep detailed records of who has visited yout Webpage, you can catch them in action. You will see them titelessly performing one of the most thankless and boring jobs humanity has ever designed: reading and indexing millions of Webpages.

Designed for speed and efficiency, these robots—the sports cars of the Web—rapidly sweep along the links, sniffing out just about everything in their paths. While these road warriors overshadowed the little beetle Hawoong Jeong built to map the Web, I was truly proud of it. It was like the first used car one could finally afford. And it crashed just

about every other day, often getting into trouble by inadvertently carrying home Webpages protected by robot exclusion riles.

It soon became clear that mapping the whole Web was a dream beyond the capabilities of our little engine. But sneaking and often stalled, it managed to carry home about 300,000 Webpages, enough to discover that there are scale-free networks out there. We shut it down at that point—perhaps a bit too early. Had we let it go further and allowed it to bring home a larger sample of the Web, we might have discovered other features of complex networks that were not so evident from our smaller sample. Search engines do see a much larger portion of the Web than we did during our experiments. Researchers studying these huge samples have made some fascinating-discoveries. They have found that the Web is fragmented into continents and communities, limiting and determining our behavior in the online universe. Paradoxically, they have also told us that there is terra incognita out there, whole continents of the Web never visited or seen by robots. Most important, we learned that the structure of the World Wide Web has an impact on everything from surfing to democracy.

1.

A few years ago we thought we knew everything there was to know about the Web. Comments like "If you can't find it using AltaVista, it's probably not out there" or "HotBot is the first search robot capable of indexing and searching the entire Web" were routine. We trusted the search engines to cover and deliver the Web to us. This suddenly changed in April 1998. "We prefer to index quality sites instead of a greater quantity of sites" was the new spin from the spokesman of a major search engine. Others went even further, claiming that "many pages are not worth indexing." What happened? This sudden mood shift was provoked by a research paper published on April 3, 1998, in the journal *Science.* Its three pages completely changed our perceptions about the accessibility of information stored on the Web.

Steve Lawrence and Lee Giles nevet planned to undermine the credibility of search engines. Working at the NEC Research Institute in Princeton, New Jetsey, they were interested in machine learning, a booming subfield of computer science. They built a meta-search engine, a robot called *Inquirus* that could inquire at each majot search engine for documents matching a given query. Halfway through they realized that their robot could do more than it was originally designed for: It could help them estimate the size of the Web.

Inquirus asked several search engines to list all documents containing a given word, for example, *crystal.* If each search engine visits and indexes the full Web, it must return the same list of documents. In reality the lists returned by different search engines are rarely identical. There is always significant ovetlap, however. For example, of the 1,000 documents containing the world crystal found by AltaVista, 343 were on HotF3ot's list as well. Dividing the number of overlapping documents by the number of documents returned by AltaVista gives the fraction of the Web covered by HotBot. Since HotBot reportedly indexed 110 million pages in December 1997, the NEC group estimated that the World Wide Web had approximately 110/0.343 million, or about 320 million documents at the same time. Today this numbet may not seem that large. In 1997, however, this was at least twice the current best guess of the Web's size.

Before 1998 we believed everything the search engines told us about the size of the Web. Aftet all, they should know. Lawrence and Giles's landmark study turned the Web into a target of scientific inquiry—one that could and must be studied using systematic and reproducible methods. But theit findings about the search engines' ability to map the Web offered us little to cheer about.

2.

Accotding to the NEC study in 1997 HotBot collected the largest number of documents, earning the distinction of being the search engine with the largest coverage. This was great news for the company. David Pritchard, marketing director for HotBot, proudly acknowledged this:

"We're the largest index out there—there are no surprises for us in this teport." Well, there were some. The bad news was that HotBot covered only 34 percent of the full Web. That is, 66 percent of all Webpages were unseen by it. AltaVista, the most populat search engine at that time, was second on the list because its robots sniffed out only 28 percent. Some search engines, such as Lycos, had captured as little as 2 percent. Their reaction was predictable: "Quite frankly, I don't give these kinds of reports a lot of credence. Our focus is not on quantity, it's on quality," said Rajive Mathur, senior product manager at Lycos Inc.

One would think that the NEC study would have motivated the search engines to increase their coverage. It didn't. A year later, in February 1999, Lawrence and Giles repeated theit measurements and found that the size of the Web had more than doubled, swelling to 800 million documents, but the search engines had not kept up with this gtowth. In fact, their covetage had futther deteriorated. This time Northern Light was the leader, covering a mere 16 percent of the Wotld Wide Web. HotBot and AltaVista had lost significant ground: Their coverage decreased to 11 and 15 percent, tespectively. Google indexed only 7.8 percent of the estimated 800 million pages out there. Taken together, in 1999 the search engines covered about 40 percent of the full Web. That means that six out of ten pages relevant to yout query would never be returned by *any* search engine. Simply, they would have nevet seen it.

Eventually the NEC study did ignite a fierce competition among the search engines. Size suddenly mattered. A fight fot dominance developed between AltaVista and the new seatch engine run by FAST, whose address, alltheweb.com, leaves little room for ambiguity regarding the company's goal. In January 2000 alltheweb.com broke the 300-million-page mark. AltaVista followed shortly. By June 2000 the new kid on the block, Google, had become a serious contendet, breaking the 500-million mark. Inktomi soon matched that, and so did yet anothet newcomer, WebTop.com. In June 2001 Google hit a new record, reaching for the first time the magic 1-billion-document covetage mark.

As of now Google maintains the lead. Alltheweb.com, pursuing its dream to eventually map out the full Web, is second with over 600

million documents, followed by AltaVista with 550 million. The search engines are doing better and better. This is great news. There is one problem, however: The Web is growing even faster.

Most search engines do not even try to reach the full Web. The reason is simple: The search engine with the most documents is not necessarily the best one. To be sure, if you are looking for difficult-to-find information, the engine with the larger coverage is your best bet. But when it comes to popular topics, a larger index does not necessarily offer better results. Most of us are already overwhelmed by the thousands of hits search engines return for simple queries. The last thing we want is to see millions more. Therefore, beyond a certain point it is more profitable to enhance the algorithm that selects the *best* page from the search engine's already enormous database than to go deeper into the Web.

When it comes to surfing the Web, either by individuals or robots, economic incentives (or their absence) are not the only limitations. The topology of the Web limits our ability to see everything out there. The World Wide Web is a scale-free network, dominated by hubs and nodes with a very large number of links. But, as we will see next, this large-scale topology coexists with numerous small-scale structures that severely limit how much we can explore simply by clicking our way along the links.

3.

Despite the billion documents on the Web, nineteen degrees of separation suggests that the Web is easily navigable. Big yet small. But the small world behind the Web is a bit misleading. To be sure, if there is a path between two documents, that path is typically short. But in reality not all pages can be connected to each other. Starting from any page, we can reach only about 24 percent of all documents. The rest are invisible to us, unreachable by surfing.

This is a consequence of the fact that for various technical reasons the links of the Web are directed. In other words, along a given URL we can travel only in one direction. If there is no direct link between

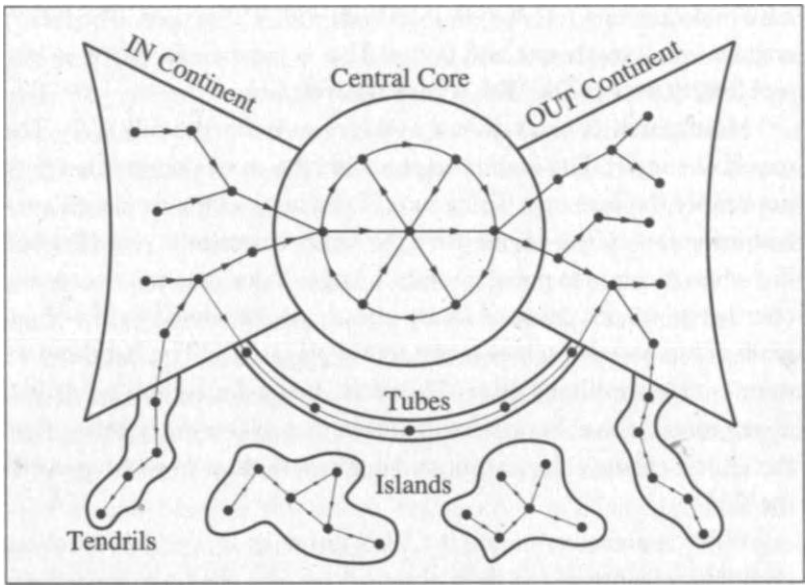

Figure 12.1 The Continents of a Directed Network. *Directed networks such as the World Wide Web naturally break down into several easily identifiable continents. In the central core each node can be reached from every other node. Nodes in the IN continent are arranged such that following the links eventually brings you back to the central core, but starting from the core doesn't allow you to return to the IN continent. In contrast, all nodes of the OUT continent can be reached from the core, but once you've arrived, there are no links taking you back. Finally, tubes directly connect the IN to the OUT continent; some nodes form tendrils, attached only to the IN and OUT continents; and a few nodes form isolated islands that can't be accessedfrom the rest of the nodes.*

two nodes in a directed network, you can connect them through other nodes: For example, if you want to go from A to D, you can start from node A, then go to node B, which has a link to node C, which points to D. But you can't make a round-trip. In a nondirected network, where you can follow a link in both directions, an A—» B —» C —» D path implies that the shortest path from D to A is the reverse one, D —» C —> B —» A. In a directed network, however, there is no guarantee that the in-

verse path exists. Most likely you would have to follow a different route back: From D you might need to visit dozens of intermediate nodes before getting back to A. The Web is full of such disjointed directed paths. They fundamentally determine the Web's navigability.

Directed networks do not represent a fundamentally new class of networks: Whether the network is scale-free or random, the links can be either directed or nondirected. So far we have dealt with mostly nondirected links. Indeed, most webs, ranging from social to protein intetaction netwotks, are nondirected. But some netwotks, ranging from the World Wide Web to food webs, have directed links. This directedness has consequences for the network's topology. In the context of the World Wide Web these consequences were first addressed by Andrei Broder, from AltaVista, and his collaborators from IBM and Compaq. They studied a sample of 200 million nodes, close to a fifth of all Webpages in existence in 1999. Theit measutements indicated that the most important consequence of ditectedness is that the Web does not form a single homogeneous network. Rathet, it is broken into four major continents (see Figure 12.1), each forcing us to obey different traffic rules when we want to navigate them.

The first of these continents contains about a quarter of all Webpages. Often called the *central core,* it gives a home to all major Websites from Yahoo! to CNN.com. Its distinguishing feature is that it is easily navigable, since there is a path between any two documents belonging to it. This does not mean that there is a direct link between any two nodes of the central core. Rather, there is a path along nodes belonging to the core that allows you to surf between any two nodes.

The second and third continents, called IN and OUT, are just as large as the centtal core but much harder to navigate. From the pages of the IN continent you can reach the central core, but there are no paths from the core taking you back to IN land. In contrast, the nodes belonging to the OUT *continent* can be easily reached from the central core, but once you have left the core, thete are no links to take you back. The OUT land is populated by corporate Websites that can be easily teached from outside; but once you get in, there is no way out. The fourth continent is made of *tendrils* and disconnected islands,

isolated groups of interlinked pages that ate unreachable from the central core and do not have links back to it. Some of these isolated groups can contain thousands of Web documents. About a quarter of all Web documents are located on such islands ot tendrils. In general the location of a Webpage on the Web has little to do with the page's content; tathet it is mostly determined by its relationship, via incoming and outgoing links, to other documents.

These four continents significantly limit the Web's navigability. How fat we can get surfing depends on where we start. Taking off from a node belonging to the central core, we can reach all pages belonging to this major continent. No matter how many times we are willing to click, however about half of the Web will still be invisible to us, since the IN land and the isolated islands cannot be reached from the core. If we step out of this core, into the OUT land, we will soon hit a dead end. If we start out journey from a tendril or an isolated island, the Web will appeat very tiny because only the othet documents on the same island will be reachable. If your Webpage is on an island, the search engines will nevet discovet it, unless you submit yout URL address to them.

Therefore, our ability to map out the full World Wide Web is not only a question of resources or economic incentives. The directedness of the links creates a very fragmented Web dominated by fout majot continents. Search engines have an easy time mapping out about half of it, the connected component and the OUT land, since the nodes belonging to them can be located starting from any node of the frequently visited central core. However, the othet half of the Web, made up of the islands and IN land, is hopelessly isolated. No matter how hatd the robots try, they will not be able to find the documents on them. This is why most search engines allow you to submit the address of your Website. If you do that, they can start crawling from it and potentially discover links to regions of the Web where they have nevet been. If you refuse to volunteer this information, many nodes could be residing in terra incognita for years to come.

Is this fragmented structure here to stay? Or will the evolving and growing Web eventually absorb the fout continents into a single, fully connected core? The answer is simple: As long as the links remain

directed, such homogenization will never occur. The continents are by no means a property peculiar to the World Wide Web. They appear in all directed networks. Consider for example a network crucial for our ability to find scientific information: the citation network. Each scientific paper cites other papers, relevant to the discussed work. A mathematics paper would cite other math papers focusing on similar problems or occasionally a biology or a physics paper, illustrating the applications of the obtained results. Therefore, all scientific publications are part of a web of science in which nodes are research publications connected by citations. These links are directed. Indeed, following the references at the end of this book will allow you to find the quoted papers. Yet none of these papers could send you to this book, since they do not cite it. The citation network is a very peculiar directed network in which the IN and OUT components reflect the historical ordering of the papers and the central component is very small if it exists at all. Nature also harbors some directed webs. In food webs, species are connected by links telling us which species feeds on which other species. The links of these networks seldom go both ways: The lion eats the antelope and never the other way around.

The bottom line is that *all* directed networks break into the same four continents. Their existence does not reflect any organizing principles particular to the Web. Random or scale-free, if the links are directed, the continents are there. This was recently demonstrated by Sergey Dorogovstev, Jose Mendes, and A. N. Samukhin, from the University of Porto, Portugal. They showed that the size and structure of these continents can be predicted analytically. Obviously, depending on the particular network's properties, the relative size of these continents varies. Yet, these results indicate that, no matter how complex and large the Web becomes, the continents are here to stay.

4.

In June 2000 Cass Sustein, a law professor at the University of Chicago, conducted a random survey of sixty political sites, finding that only 15 percent of them have links to sites with opposite views.

In contrast, as many as 60 percent have links to like-minded Webpages. A study focusing on democratic discourse on the Web arrived at a similar conclusion: Only about 15 percent of Webpages offer links to opposing viewpoints. Sustein fears that by limiting access to conflicting viewpoints, the emerging online universe encourages segregation and social fragmentation. Indeed, the mechanisms behind social and political isolation on the Web are self-reinforcing: They alter the Web's topology as well, segregating the online universe. Therefore, the four continents are not the only isolated structures of the Web. On a smaller scale, these continents are sprinkled with vibrant villages and metropoli. These are Websites brought together by a joint idea, hobby, or habitat, forming communities of shared interests. Jazz enthusiasts form a well-defined Web-based community, but so do bird-watchers. Religious fundamentalists in eastern Europe share virtual space with their ideological counterparts in the United States. Antiglobalization activists in Europe join forces with their peers in Japan to coordinate strategies and activities.

Communities are essential components of human social history. Granovetter's circles of friends, the elementary building blocks of communities, pointed to this fact. Lately, however, perhaps unrecognized by their members, such communities are increasingly recorded in the Web's topology. A side effect of our digital life is that our beliefs and affiliations are publicly available. Each time we link to a Webpage, we are endorsing its relevance to our intellectual curiosity. Thus the links of an enthusiastic bird-watcher can take us to other like-minded Web sites, allowing us to map out the community of bird enthusiasts.

Identifying such Web-based communities has tremendous potential for applications. Indeed, finding the community of sports car enthusiasts would allow car companies to most effectively market their new-models by placing ads at several hubs of this community. AIDS activists could use community knowledge to mobilize those who passionately care about the disease, molding them into an effective lobbying and action group. Organizers of ethnic festivals could take advantage of information about Web-based ethnic communities to advertise upcoming events and incubate local grassroots organizations. The problem is that

there are a billion-plus pages out there- Can we locate communities on such a gigantic web?

Supreme Court justice Potter Stewart famously remarked in 1964 that "I shall not today attempt further to define [obscenity]... and perhaps I never could succeed in intelligibly doing so. But I know it when I see it." We face similar problems when we try to find a proper definition of "Web-based communities." We all know them once we see them, but everybody has slightly different criteria for them. One reason is that there are no sharp boundaries between various communities. Indeed, the same Website can belong simultaneously to different groups. For example, a physicist's Webpage might mix links to physics, music, and mountain climbing, combining professional interests with hobbies. In which community should we place such a page? The size of communities also varies a lot. Fot example, while the community interested in "cryptography" is small and relatively easy to locate, the one consisting of devotees of "English litetature" is much harder to identify and fragmented into many subcommunities ranging from Shakespeare enthusiasts to Kurt Vonnegut fans.

Recently Gary Flake, Steve Lawrence, and Lee Giles, from NEC, suggested that documents belong to the same community if they have more links to each othet than to documents outside of the community. This definition is precise enough to develop algorithms to identify different groupings given the topology of the World Wide Web. It turns out, however, that actually finding these communities is notoriously difficult. This kind of search belongs to the class of so-called NP complete problems, which means that, though in principle communities can be located, there is no efficient algorithm for doing so. Therefore, the difficulty in finding communities on the Web is similar to solving the traveling salesperson's problem, which asks us to find the shortest route reaching a given number of cities assuming that we are not allowed to visit the same city twice. The only algorithm guaranteed to work for finding communities or the route for the traveling salesperson requires us to try all possible combinations. For communities, the time required to perform such a search increases exponentially with the size of the Web. With fast enough computers we might be able to locate communities in a sample of

a hundred documents. Uncovering them from a billion Webpages, however, is simply out of the question.

Combining content and topology makes the problem somewhat less challenging. For example, we can focus on documents that contain only one or two keywords. Lada Adamic, from Stanford University, recently investigated communities discovered by searching for the phrases "abortion—pro choice" and "abortion—pro life." The pro-life query resulted in a core of forty-one documents in which you could get from each page to the other ones. In contrast, the pro-choice movement was fragmented into many disconnected sites.

Such differences in the structure of competing communities have important consequences for their ability to market and organize themselves for a common cause. As Adamic notes, a campaign against the partial birth abortion bill launched from the middle of the pro-life cluster could easily reach other pro-life sites, since there are many links between them. Furthermore, due to the links on the pro-choice sites, the visitors of pro-choice sites would also learn about it. However, one would need to advertise at several disconnected pro-choice Websites to achieve an equally efficient campaign against the bill. Therefore, not only does the pro-life community have a better presence on the Web, it is also better organized—its sites are more aware of each other.

Far from being a homogenous sea of nodes and links, the Web is fragmented into four continents, each of which hosts many villages and cities that appear as overlapping communities. Any of us willing to take up a virtual presence belongs to one or several of them. To be sure, we are far from fully understanding this fine structure of the Web. But many forces, from commercial interests to scientific curiosity, increasingly motivate us to do better. As we dig deeper, I am sure that we will encounter many surprises, offering us an even clearer view of this complex, amorphous, ever changing online universe.

5.

On November 20, 2000, in a precedent-setting decision, Judge Jean-Jacques Gomes of France ordered Yahoo! to deny French consumers

access to any of its sites that auction Nazi memorabilia. It did so by upholding a French law prohibiting the sale of such items in France. The legal implications of the court's decision are still being debated across the world. Yahoo! argued that the Internet is fundamentally free from geographic and national boundaries and that subjecting the U.S. company to national laws around the world was therefore a severe breach of the Internet's basic philosophy. Others disagreed, saying that there is nothing particularly novel about the Internet and that it should be covered by the same international trade agreements as any international business.

Beyond the legal tamifications, the deeper issue is about the code—the software behind the Web. The Ftench coutt acknowledged that considering the nature of the Web, there is no way to keep France completely isolated from the world. They were persuaded, however, by experts who testified that Yahoo! could put in place a filtering mechanism that would block at least 70 to 80 percent of French nationals trying to reach Yahool's Nazi sites. Thus, the court otdered Yahoo! to alter the code. This is exactly the type of action that Lawrence Lessig, a Stanford University law professor, envisioned in his influential book Code *and Other Laws of Cyberspace.* According to Lessig, "Left to itself, cyberspace will become a perfect tool of control.... [T]he invisible hand of cyberspace is building an architecture that is quite the opposite of what it was at the cyberspace's birth."

Lessig uses the word *architecture* to mean the sum of all softwate running behind the Web, concluding that the only way to influence behavior in cyberspace is to regulate the code. He suggests that two forces are aligned to do just that. First, governments have a hard time policing behavior on the Web. It is easy to write legislation limiting access to everything from pornography to keys to cryptographic codes. In a borderless cyberworld, however, it is almost impossible to enforce these laws. If governments pass on the opportunity to regulate the Web, commetce will live with it. Companies seeking a more secure business environment in which they can identify customers for various purposes ranging from security concerns to marketing will push the code in the direction of control. Netizens will completely lose theit

anonymous and space-free existence as the technology develops to meet the merchants' desires.

On one hand, as the Yahoo! case and others have demonstrated, some of Lessig's bleak predictions have become teality. On the other hand, in my view, to truly understand cyberspace we need to distinguish carefully between *code* and architecture. Code—or software—is the bricks and mortar of cyberspace. The architecture is what we build, using the code as building blocks. The great architects of human history, from Michelangelo to Frank Lloyd Wright, demonstrated that, whereas raw materials are limited, the architectural possibilities are not. Code can curtail behaviot, and it does influence the architecture. It does not uniquely determine it, however.

Like architects' buildings, the Web's architecture is the product of two equally important layers: *code* and *collective human actions* taking advantage of the code. The first can be regulated by courts, government, and companies alike. The second, however, cannot be shaped by any single user or institution, because the Web has no central design—it is self-organized. It evolves from the individual actions of millions of users. As a result, its architectute is much richer than the sum of its parts. Most of the Web's truly important features and emerging properties derive from its large-scale self-organized topology.

A good example is democracy on the Web. We've seen that the scale-free topology means that the vast majority of documents ate hardly visible, since a highly popular minority has all the links. Yes, we do have free speech on the Web. Chances ate, however, that out voices are too weak to be heard. Pages with only a few incoming links are impossible to find by casual browsing. Instead, over and over we are steered towatd the hubs. It is tempting to believe that robots can avoid this popularity-driven trap. They could, but they don't. Instead, the likelihood that a document will be indexed by a seatch engine depends strongly on the numbet of its incoming links. Documents with only one incoming link have less than a 10 percent chance of being noticed by any search engine. In conttast, robots find and index close to 90 percent of pages that have twenty-one to one hundred incoming links.

Lessig is right: The architecture of the Web controls just about everything, from access to consumers to the probability of being visited by surfing along the links. But the science of the Web increasingly proves that this architecture represents a higher level of organization than the code. Your ability to find my Webpage is determined by one factor only: its position on the Web. If many people find my page interesting and they link to me, my node will slowly turn into a minor hub, and search engines will inevitably notice. If everybody ignores my Website, so will the search engines. I will join the ranks of invisible Websites, which are the majority anyway. Thus the Web's large-scale topology—that is, its true architecture—enforces more severe limitations on our behavior and visibility on the Web than government or industry could ever achieve by tinkering with the code. Regulations come and go, but the topology and the fundamental natural laws governing it are time invariant. As long as we continue to delegate to the individual the choice of where to link, we will not be able to significantly alter the Web's large-scale topology, and we will have to live with the consequences.

6.

The great thing about the Web is that our Webpages mature with us. Once we alter our personal page, nobody can haunt us with the opposite views we might have held decades earlier. Do you remember that boyfriend you broke up with a few years ago? Of course you do, but you probably hope that nobody else does. To be sure, all his pictures are gone from your Webpage. How about that high-school manifesto you are still embarrassed about? Or that collection of links to Democratic sites you assembled a mere two years before running on a Republican ticket? They are all untraceable. Or at least, we tend to think so. That is because most netizens have never heard of Brewster Kahle. The truth is, Kahle could easily have a copy of all the pictures and documents that you so carefully removed from your Website and have now forgotten.

The inventor of wide area information servers and founder of Alexa Internet, one of the major search engines, Kahle is a veteran ^ **of** the Web. After selling Alexa to Amazon.com in 1999, he used the

proceeds to create the Internet Archives, a nonprofit organization located in Presidio, a converted military base in downtown San Francisco. His goal is simple: He wants to prevent the Web's content from disappearing into the past.

When I visited the Archives to give a talk at the First Internet Archive Workshop in March 2000, Kahle reminded me of the ancient library of Alexandria. It was believed to have had a copy of all books written in the ancient word, all of which disappeared when the library was burned to the ground. He also told me about great cinematographic collections that were recycled for their silver content. Without cultural artifacts, humanity has no memory, and without memory it cannot learn from its successes and failures. When it comes to the,World Wide Web, we are again letting history go unrecorded. To avoid repeating history, Kahle's brainchild, the Internet Archives, carefully keeps all documents that Alexa has crawled to since 1996. The collection has already swelled to 100 billion Webpages, representing about 100 terabytes of information. In comparison, all books and documents archived by the Library of Congress are only about 20 terabytes.

The Archives' collection is of unparalleled value for historians, social scientists, and Web topographers alike. To write the history of the 2000 presidential election, you would start with the Archives. They have a time machine that allows you to see the candidate's sites, voter guides, and the Web pages of political parties, exactly as they were during the campaign. Do you want to track the reaction of the online universe to the September 11, 2001, terrorist attacks? One month after the events the Archives already had a collection of 200 million related documents. If you are a Web topographer aiming to understand the Web's architecture, the Archives are an excellent starting point. They let you trace when and where Webpages and links were added and removed, how some latecomer nodes become popular overnight; and how former hubs lost their shine. Comparing the maps of the Web taken at different time intervals, you can follow the emergence and crystallization of virtual communities. The Archives have the data to reconstruct the chaotic evolution of nodes and links, helping to uncover the mechanisms responsible for the Web's current architecture.

The Archives have many fans from many different disciplines, but most tesearchers who could take advantage of them either do not know of theit existence or lack the programming skills to access and efficiently use them. So their full potential is still untapped by researchers and the public alike. Howevet, I hope that the Internet Archives are only the beginning of an awakening to our historical responsibility towatds the online universe. The Archives are far from capturing everything out there. Their main collection comes from Alexa, the search engine founded by Kahle and Bruce Gilliat in 1996. As we already know, search engines covet only a small fraction of the World Wide Web, and Alexa was never known for pursuing a significant coverage. Therefore, despite their enormous size, the Archives' current collection represents only a tiny fraction of the Web, mostly popular Webpages. Alexa got the hubs; the rest, the vast majority of less connected pages ignored by their robots, are slipping into oblivion at a rate of millions per day.

7.

To an alien approaching our solar system, the Earth would appear to be nothing more than a spherical ball. Getting closer, the alien might start noticing the continents. The bright lights of Paris, New York City, London, and Tokyo offer clues of intelligent life. Getting even closer, smaller communities become discernable, and a fine structure of connecting highways and roads emerges. The alien would have to come really close, however, to see the human beings responsible for the large-scale oider visible from space.

Our exploration of the Wotld Wide Web has followed an identical route. First we discovered the inhomogeneous large-scale topology and understood that it is as unavoidable as the spherical shape of most planets. Looking closet, we noticed four major continents, each obeying different laws. Btinging more details into focus, we started to see communities, groups of Webpages held together by common interests. These forays into the unknown have significantly altered our understanding of the World Wide Web. We learned that the online universe is much larger than anyone ever anticipated. It also grows fastet than we were

ready to believe. To our dismay, we also found that it is much less chatted than we were willing to accept. Two years ago, six out of ten pages had not been visited by any search engine. If the trend can be trusted, today's search engines see an even smaller fraction of the Web. The good news is that competition forces the search engines to do a better job. But we should never lose sight of the big picture: Whatevet the extent of their competition, the Web is even bigger.

Yet we shouldn't underestimate the enormous services the search engines and their robots offer us. We often sigh in desperation, calling the Web a "jungle." The truth is, without tobots it would be a black hole. Space would curve around it such that anything falling in would never get out. Robots keep the Wotld Wide Web from collapsing under its increasing complexity. They fold the space out, maintaining otdet in the chaos of nodes and links.

Our life is increasingly dominated by the Web. Yet we devote remarkably little attention and resources to understanding it. Relatively little effort would be required to bring along a new revolution in information access. It will happen. The question is, what do we lose in the meantime?

In an increasingly Internet-dominated society, understanding the World Wide Web has tremendous value in and of itself. For me, however, the rewards go beyond that. One of the most exciting aspects of this explotation has been uncovering laws whose validity does not stop at the gates of cyberspace. These laws, applying equally well to the cell and the ecosystem, demonstrate how unavoidable nature's laws are and how deeply self-organization shapes the world around us. By virtue of its digital nature and enormous size the World Wide Web offers a model system whose every detail can be uncovered. We have never gotten this close to any network before. It will continue to be a source of inspiration and ideas to anybody aiming to grasp the properties of our weblike universe.