Web Science EMERGES

Studying the Web will reveal better ways to exploit information, prevent identity theft, revolutionize industry and manage our ever growing online lives

By Nigel Shadbolt and Tim Berners-Lee

KEY CONCEPTS

- The relentless rise in Web pages and links is creating emergent properties, from social networking to virtual identity theft, that are transforming society.
- A new discipline, Web science, aims to discover how Web traits arise and how they can be harnessed or held in check to benefit society.
- Important advances are beginning to be made; more work can solve major issues such as securing privacy and conveying trust.

—The Editors

Since the World Wide Web blossomed in the mid-1990s, it has exploded to more than 15 billion pages that touch almost all aspects of modern life. Today more and more people's jobs depend on the Web. Media, banking and health care are being revolutionized by it. And governments are even considering how to run their countries with it. Little appreciated, however, is the fact that the Web is more than the sum of its pages. Vast emergent properties have arisen that are transforming society. E-mail led to instant messaging, which has led to social networks such as Facebook. The transfer of documents led to file-sharing sites such as Napster, which have led to user-generated portals such as YouTube. And tagging content with labels is creating online communities that share everything from concert news to parenting tips.

But few investigators are studying how such emergent properties have actually happened, how we might harness them, what new phenomena may be coming or what any of this might mean for humankind. A new branch of science—Web science—aims to address such issues. The timing fits history: computers were built first, and computer science followed, which subsequently improved computing significantly. Web science was launched as a formal discipline in November 2006, when the two of us and our colleagues at the Massachusetts Institute of Technology and the University of Southampton in England announced the beginning of a Web Science Research Initiative. Leading researchers from 16 of the world's top universities have since expanded on that effort.

This new discipline will model the Web's structure, articulate the architectural principles that have fueled its phenomenal growth, and discover how online human interactions are driven by and can change social conventions. It will elucidate the principles that can ensure that the network continues to grow productively and settle complex issues such as privacy protection and intellectual-property rights. To achieve these ends, Web science will draw on mathematics, physics, computer science, psychology, ecology, sociology, law, political science, economics, and more.

Of course, we cannot predict what this nascent endeavor might reveal. Yet Web science has already generated crucial insights, some presented here. Ultimately, the pursuit aims to answer fundamental questions: What evolutionary patterns have driven the Web's growth? Could they burn out? How do tipping points arise, and can that be altered?

Insights Already

Although Web science as a discipline is new, earlier research has revealed the potential value of such work. As the 1990s progressed, searching for information by looking for key words among the mounting number of pages was returning more and more irrelevant content. The founders of Google, Larry Page and Sergey Brin, realized they needed to prioritize the results.

Their big insight was that the importance of a page—how relevant it is—was best understood in terms of the number and importance of the pages linking to it. The difficulty was that part of this definition is recursive: the importance of a page is determined by the importance of the

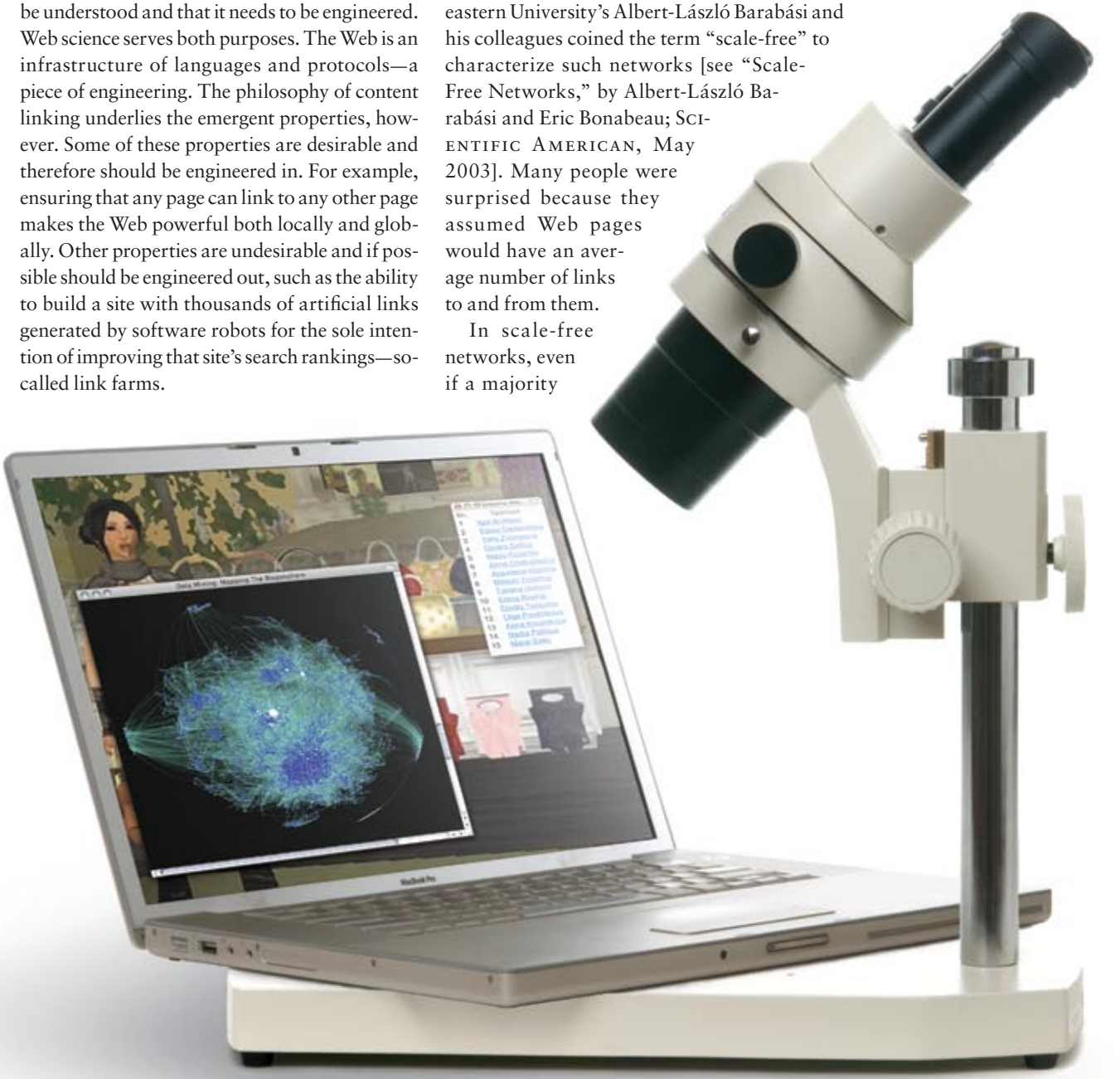
CARY WOLINSKY; MATTHEW HURST *Microsoft Live Labs (http://datamining.typepad.com/data_mining) (blogosphere visualization)*

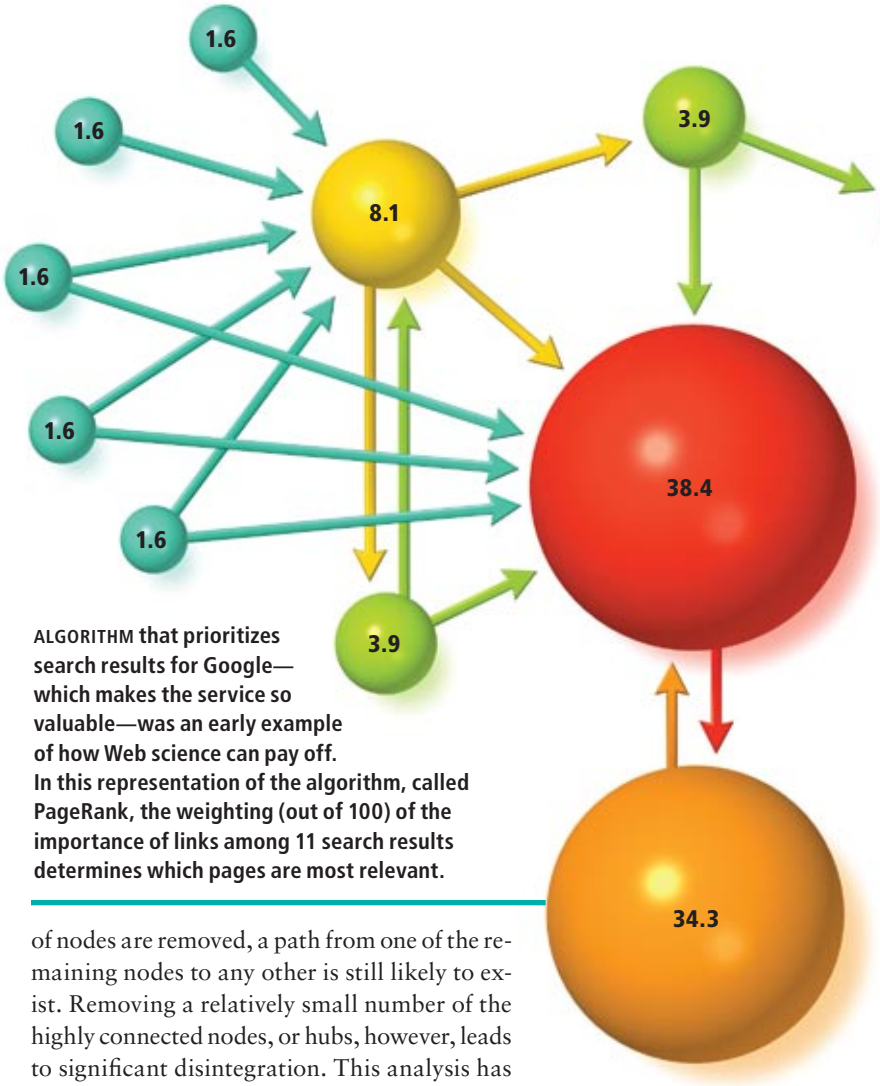
pages linking to it, whose importance is determined by the importance of the pages linking to them. Page and Brin figured out an elegant mathematical way to represent that property and developed an algorithm they called PageRank to exploit the recursiveness, thus returning pages ranked from most relevant to least.

Google's success shows that the Web needs to be understood and that it needs to be engineered. Web science serves both purposes. The Web is an infrastructure of languages and protocols—a piece of engineering. The philosophy of content linking underlies the emergent properties, however. Some of these properties are desirable and therefore should be engineered in. For example, ensuring that any page can link to any other page makes the Web powerful both locally and globally. Other properties are undesirable and if possible should be engineered out, such as the ability to build a site with thousands of artificial links generated by software robots for the sole intention of improving that site's search rankings—so-called link farms.

Another early discovery, which came from graph theory, is that the Web's connectivity follows a so-called power-law degree distribution. In many networks, nodes have about the same number of links to them. But on the Web a few pages have a huge number of other pages linking to them, and a very large number of pages have only a few pages linking to them. Northeastern University's Albert-László Barabási and his colleagues coined the term "scale-free" to characterize such networks [see "Scale-Free Networks," by Albert-László Barabási and Eric Bonabeau; SCIENTIFIC AMERICAN, May 2003]. Many people were surprised because they assumed Web pages would have an average number of links to and from them.

In scale-free networks, even if a majority





ALGORITHM that prioritizes search results for Google—which makes the service so valuable—was an early example of how Web science can pay off. In this representation of the algorithm, called PageRank, the weighting (out of 100) of the importance of links among 11 search results determines which pages are most relevant.

of nodes are removed, a path from one of the remaining nodes to any other is still likely to exist. Removing a relatively small number of the highly connected nodes, or hubs, however, leads to significant disintegration. This analysis has been crucial for the companies and organizations—be they telecommunications providers or research laboratories—that design how information is routed on the Web, allowing them to build in substantial redundancy that balances traffic and makes the network more resistant to attack.

Thorough understanding of scale-free networks, gleaned by analyzing the Web, has prompted experts to analyze other network systems. They have since found power-law degree distributions in areas as far-flung as scientific citations and business alliances. The work has helped the U.S. Centers for Disease Control and Prevention improve its models of sexual disease transmission and has helped biologists better understand protein interactions.

Scientific analysis has also characterized the Web as having short paths and small worlds. While at Cornell University in the 1990s, Duncan J. Watts and Steven H. Strogatz showed that even though the Web was huge, a user could get from one page to any other page in at most 14 clicks. Yet to fully understand these traits, we need to appreciate that the Web is a social network. In 1967 Harvard University psychologist Stanley Milgram asked residents in Omaha, Neb., and Wichita, Kan., to attempt to send a package to an individual described only by his name, some general features and the fact that he lived in Boston. The residents were to send the package to an intermediary who they thought might know more about how to reach the individual and who could then send it on to another intermediary. Eventually 64 of the almost 300 packages made it to the designated recipients. On average the number of intermediaries needed was six—the basis of the catchphrase "six degrees of separation."

More recently, however, Watts, now at Columbia University, tried to repeat the experiment with an e-mail message to be forwarded on the Web and experienced failures in path finding. In particular, if individuals had no incentive to forward the note the paths broke down. Yet only very slight incentives improved matters.

The lesson is that network structure alone is not everything; networks thrive only in the light of the actions, strategies and perceptions of the individuals embedded in them. To realistically know why the Web has a beneficial structure of short paths, we need to know why people who contribute content link it to other material. Social drivers—goals, desires, interests and attitudes—are fundamental aspects of how links are made. Understanding the Web requires insights from sociology and psychology every bit as much as from mathematics and computer science.

From Micro to Macro

One major area of Web science will explore how a small technical innovation can launch a large social phenomenon. A striking example is the emergence of the blogosphere. Although early Web browsers did not provide a handy way for the average person to "publish" his or her ideas, by 1999 blog programs had made self-publishing much easier. Blogging subsequently caught fire because as people got issues off their chest, they also found others with similar views who could readily assemble into a like-minded community.

It is difficult to estimate the size of the blogosphere accurately. David Sifry's leading blog search engine, called Technorati, was tracking more than 112 million blogs worldwide in May of this year, a number that may include only a mere fraction of the 72 million blogs purportedly in China. Whatever the size, the explosive

SEXUAL ORIENTATION REVEALED

Carter Jernigan and Behram Mistree, both students at the Massachusetts Institute of Technology, recently analyzed Facebook communities. They found that the structure identifies the likely sexual orientation of individuals who have not explicitly stated their preference. The reason is that people who have made a declaration link more often to others of a similar orientation; a kind of triangulation emerges. This sort of result raises ethical and privacy questions; more research about the Web's structure and users' behavior could provide answers.

GEORGE RETSECK

growth demands an explanation. Arguably, the introduction of very simple mechanisms, especially TrackBack, facilitated the growth. If a blogger writes an entry commenting on or referring to an entry at another blog, TrackBack notifies the original blog with a "ping." This notification enables the original blog to display summaries of all the comments and links to them. In this way, conversations arise spanning several blogs and rapidly form networks of individuals interested in particular themes. And here again large portions of the blog structure become linked via short paths—not only the blogs and bloggers themselves but also the topics and entries made.

As blogging blossomed, researchers quickly created interesting tools, measurement techniques and data sets to try to track the dissemination of a topic through blogspace. Social-media analyst Matthew Hurst of Microsoft Live Labs collected link data for six weeks and produced a plot of the most active and interconnected parts of the blogosphere [*see illustration on next page*]. It showed that a number of blogs are massively popular, seen by 500,000 different individuals a day. A link or mention of another blog by one of these superblogs guarantees a lot of traffic to the site referenced. The plot also shows isolated groups of dedicated enthusiasts who are very much in touch with one another but barely connect to other bloggers.

If exploited correctly, the blogosphere can be a powerful medium for spreading an idea or gauging the impact of a political initiative or the likely success of a product launch. The much anticipated release of the Apple iPhone generated 1.4 percent of all new postings on its launch day. One challenge is to understand how this dissemination might change our view of journalism and commentary. What mechanisms can assure blog readers that the facts quoted are trustworthy? Web science can provide ways to check this so-called provenance of information, while offering practical rules about conditions surrounding its reuse. Daniel Weitzner's Transparent Accountable Datamining Initiative at M.I.T. is doing just that.

Rise of Semantics

One emerging phenomenon that is benefiting from concerted research is the rise of the Semantic Web—a network of data on the Web. Among many payoffs, the Semantic Web promises to give much more targeted answers to our questions. Today searching Google for "Toyota used cars for sale in western Massachusetts under $8,000" returns more than 2,000 general Web pages. Once Semantic Web capabilities are added, a person will instead receive detailed information on seven or eight specific cars, including their price, color, mileage, condition and owner, and how to buy them.

Engineers have devised powerful foundations for the Semantic Web, notably the primary language—the Resource Description Framework (RDF)—which is layered on top of the basic HTML and other protocols that form Web pages. RDF gives meaning to data through sets of "triples." Each triple resembles the subject, verb and object of a sentence. For example, a triple can assert that "person X" [subject] "is a sister of" [verb] "person Y" [object]. A series of triples can determine that [car X] [is brand] [Toyota]; that [car X] [condition is] [used]; that [car X] [costs] [$7,500]; that [car X] [is located in] [Lenox]; and that [Lenox] [is located in] [western Massachusetts]. Together these triples can conclude that car X is indeed a proper answer to our query. This simple triple structure turns out to be a natural way to describe a large majority of the data processed by machines. The subjects, verbs and objects are each identified by a Universal Resource Identifier (URI)—an address just like that used for Web pages. Thus, anyone can define a new concept, or a new verb, by defining a URI for it on the Web.

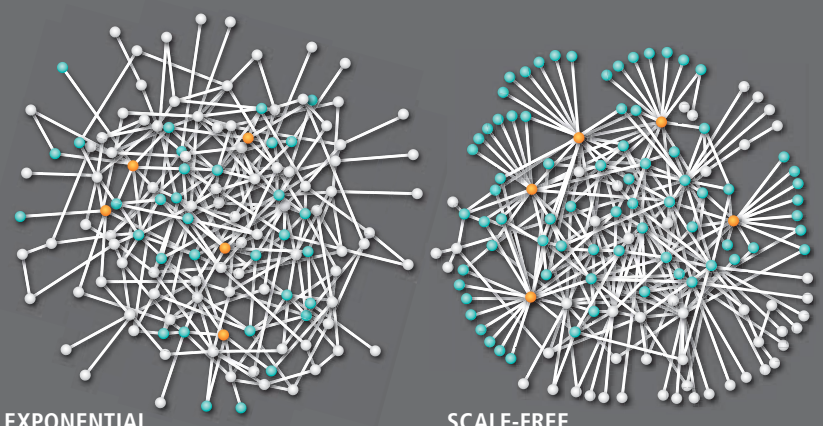
[THE AUTHORS]

Nigel Shadbolt (*left*) **is professor of artificial intelligence at the University of Southampton in England, chief technology officer of the Semantic Web company Garlik Ltd., and a past president of the British Computer Society. Tim Berners-Lee** (*right*) **invented the World Wide Web and leads the World Wide Web Consortium, based at the Massachusetts Institute of Technology.**

[WORK IN PROGRESS]

Building a More Secure Web

Understanding how the Web is linked can reveal ways to better engineer it. Many networks are fairly homogeneous (an "exponential" structure): nodes, even the busiest (*orange*) and their immediate neighbors (*blue*), have roughly the same number of links to and from them. But analysis at the University of Notre Dame showed that the Web is a scale-free network: a few nodes (Web sites) have many links coming in, and many nodes have only a few links.



JASON BUCKNER (*Shadbolt*); DONNA COVENEY (*Berners-Lee*); SOURCE: "ERROR AND ATTACK TOLERANCE OF COMPLEX NETWORKS," BY RÉKA ALBERT, HAWOONG JEONG, ALBERT-LÁSZLÓ BARABÁSI ET AL., IN *NATURE*, VOL. 406; JULY 27, 2000 (*scale-free network*)

MATTHEW HURST *Microsoft Live Labs (http://datamining.typepad.com/data_mining) (blogosphere visualization)*

BLOGOSPHERE has certain patterns of power. Matthew Hurst tracked how blogs link to one another. A visualization of the result (*left*) displays each blog as a white dot; the few large dots are massively popular sites. Blogs that share numerous cross citations form distinct communities (*purple*). Isolated groups that communicate frequently among themselves but rarely with others appear as straight lines along the outer edges.

As these definitions grow and interlink, specialists and enthusiasts will define taxonomies and ontologies: data sets that describe classes of objects and relations among them. These sets will help computers everywhere to find, understand and present targeted information.

Numerous groups are already building Semantic Web frameworks, especially in biology and health care [see "The Semantic Web in Action," by Lee Feigenbaum; SCIENTIFIC AMERICAN, December 2007]. More than 1,000 people attended the Semantic Technology conference in San Jose, Calif., this past May. Web science offers the prospect of creating more powerful ways to define, link and interpret data.

The wiki world offers a good example of how useful such exploitation of linked data can be. As of May, Wikipedia, the online encyclopedia generated by people everywhere, had more than 2.3 million articles in English. The articles contain regular text, along with infobox templates—sets of facts. More than 700,000 English infobox templates now exist, and programmers are looking for ways to mine them. One effort is the DBpedia project, begun by Chris Bizer and his colleagues at the Free University of Berlin and the University of Leipzig in Germany. They have devised a tool by the same name (available at http://wikipedia.aksw.org) that uses Semantic Web techniques to query the infoboxes. It can ask for all tennis players who live in Moscow or the names of all the mayors of towns in the U.S. that are at an altitude greater than 1,000 meters [*see box below*] and get back an exact answer.

Naturally, we would like a similar tool for the entire Web, but developing one would require that more and more data on the Web were represented as linked sets of RDF. Meanwhile it is becoming apparent that DBpedia's link structure obeys the same power laws that have been found for the Web. Just as some pages have a higher rank in the Web of documents, so it will be for data on the Semantic Web. At the same time, research by Oded Nov of the Polytechnic Institute of New York University is beginning to ascertain why Wikipedians post entries and what motivates their activity; the psychological drivers that are revealed will help us understand how to encourage people to contribute to the Semantic Web.

Future Challenges

It seems sensible to say that Web science can help us engineer a better Web. Of course, we do not fully know what Web science is, so part of the new discipline should be to find the most powerful concepts that will help the science

[CASE STUDY]

Tennis, Anyone?

Web science is generating tools that can understand online data (known as the Semantic Web) and can therefore provide highly targeted search results. One effort, DBpedia, can assess the information in infoboxes on Wikipedia pages. For example, to find all tennis players from Moscow, a user fills in a simple form (*below*) and gets a short, exact list as a result (*right*).

Subject	Predicate	Object
person	placebirth	Moscow
person	type	tennis_players

Semantic Wiki...

Nr.	?person
1	Igor Andreev
2	Elena Dementieva
3	Vera Zvonareva
4	Dinara Safina
5	Maria Kirilenko
6	Anna Chakvetadze
7	Anastasia Myskina
8	Mikhail Youzhny
9	Tatiana Golovin
10	Elena Bovina
11	Dmitry Tursunov
12	Olga Poutchkova
13	Anna Kournikova
14	Nadia Petrova
15	Marat Safin

SECOND LIFE and other virtual worlds are constructed by many enthusiasts who each contribute small pieces, such as stores and products that constitute a virtual mall where real people can profit from online avatars that shop there. Such creations raise complicated questions about who owns what intellectual property—questions Web science hopes to answer.

itself grow. Perhaps insights will come from the work's interdisciplinary nature. For example, biological concepts such as plasticity might prove useful. The brain and nervous system grow and adapt over our lifetimes by forming and deleting connections between neurons—the brain cells that act as nodes in our neural network. Changes in the connections occur in response to activity in the network, including learning, disuse and aging.

Similarly, Web connections decay and grow. Web science could also explore the possibility of protocols that disconnect Web nodes if there is no inbound or outbound activity. Would such a network function more effectively?

Concepts such as population dynamics, food chains, and consumers and producers all have counterparts on the Web. Perhaps methods and models devised for ecology can help us understand the Web's digital ecosystem, which could be prone to damage by single major events (analogous to hurricanes) or subtle but steady erosions (like invasive species).

We will also need to examine a range of legal issues. Laws relating to intellectual property and copyright for digital material are already being debated. Fascinating issues have arisen in virtual environments such as Second Life; for example, do laws and entitlements transfer to digital worlds, where millions of people contribute tiny additions to existing content? Another issue is whether we can build rules of use into content itself. An example of such a framework, called Creative Commons, lets authors, scientists, artists and educators easily mark their creative work with the freedoms and restrictions they want it to carry. Crucially, the mark also provides RDF data that describe the license, making it easy to automatically locate works and understand their conditions of use. Web science could determine whether commons-style licenses affect the spread of the information.

MORE TO EXPLORE

Exploring Complex Networks. Steven H. Strogatz in *Nature,* Vol. 410, pages 268–276; March 8, 2001.

The Semantic Web Revisited. Nigel Shadbolt, Tim Berners-Lee and Wendy T. Hall in *IEEE Intelligent Systems,* Vol. 21, No. 3, pages 96–101; May/June 2006.

Creating a Science of the Web. Tim Berners-Lee, Wendy T. Hall, James W. Hendler, Nigel Shadbolt and Daniel Weitzner in *Science,* Vol. 313, pages 769–771; August 11, 2006.

Google's PageRank and Beyond: The Science of Search Engine Rankings. Amy N. Langville and Carl D. Meyer. Princeton University Press, 2006.

Web Science: An Interdisciplinary Approach to Understanding the Web. James Hendler, Nigel Shadbolt, Wendy T. Hall, Tim Berners-Lee and Daniel Weitzner in *Communications of the ACM,* Vol. 51, No. 1, pages 60–69; July 2008.

Sociology is another field to tap. Research is needed, for instance, to provide Web users with better ways of determining whether material on a site can be trusted. How can we determine whether we can trust the material emanating from a site? The Web was originally conceived as a tool for researchers who trusted one another implicitly; strong models of security were not built in. We have been living with the consequences ever since.

As a result, substantial research should be devoted to engineering layers of trust and provenance into Web interactions. The coming together of our digital and physical personas presents opportunities for progress, such as the integration of financial, medical, social and educational services for each of us. But it is also an opportunity for identity theft, cyberstalking and cyberbullying, and digital espionage. Web science can help enhance the good and ameliorate the bad.

Various other questions need to be tackled before the rich potential of the Web can be mined to its fullest. How do social norms influence emerging capabilities? How can online privacy protection, intellectual-property rights and security be implemented? What trends could fragment the Web?

Many people are working on parts of these questions. Web science can bring their efforts together and compound the insights. We need to train a cadre of researchers, developers, practitioners and users in a broad range of skills and subjects. They will help us fully understand the Web and discover how to engineer it for the 21st century and beyond. ■

LINDEN LAB