DOI: 10.1145/1364782.1364798

The Web must be studied as an entity in its own right to ensure it keeps flourishing and prevent unanticipated social effects.

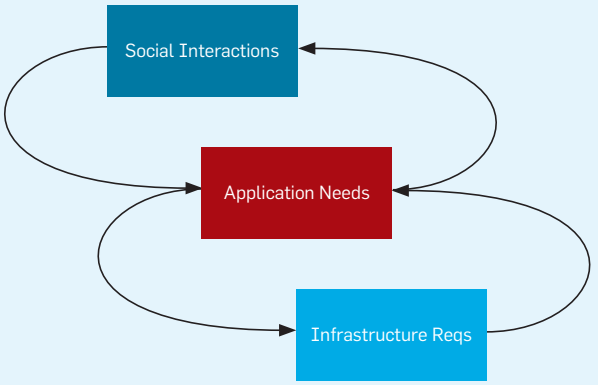
BY JAMES HENDLER, NIGEL SHADBOLT, WENDY HALL, TIM BERNERS-LEE, AND DANIEL WEITZNER

Web Science: An Interdisciplinary Approach to Understanding the Web

DESPITE THE WEB'S great success as a technology and the significant amount of computing infrastructure on which it is built, it remains, as an entity, surprisingly unstudied. Here, we look at some of the technical and social challenges that must be overcome to model the Web as a whole, keep it growing, and understand its continuing social impact. A systems approach, in the sense of "systems biology," is needed if we are to be able to understand and engineer the future Web.

ILLUSTRATION BY MARIUS WATZ

Figure 1: The social interactions enabled by the Web put demands on the Web applications behind them, in turn putting further demands on the Web's infrastructure.



Despite the huge effect the Web has had on computing, as well as on the overall field of computer science, the best keyword indicator one can find in the ACM taxonomy, the one by which the field organizes many of its research papers and conferences, is "miscellaneous." Similarly, if you look at CS curricula in most universities worldwide you will find "Web design" is taught as a service course, along with, perhaps, a course on Web scripting languages. You are unlikely to find a course that teaches Web architecture or protocols. It is as if the Web, at least below the browser, simply does not exist. Many "information schools" and "informatics departments" offer courses that focus on applications on the Web or on such topics as "Web 2.0," but the protocols, architectures, and underlying principles of the Web per se are rarely covered.

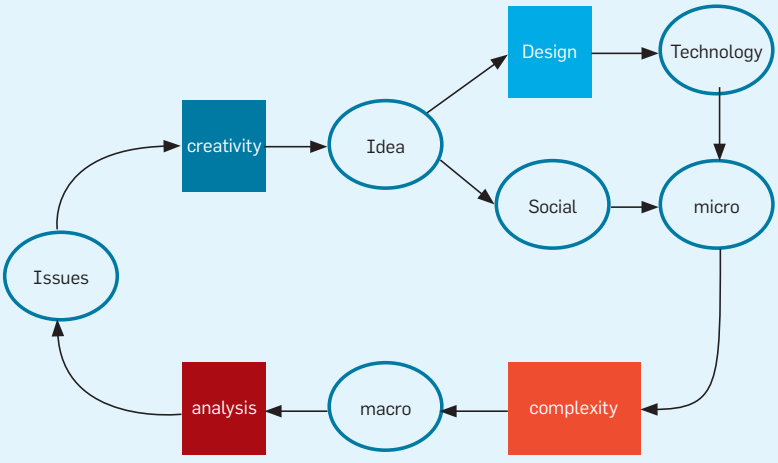
Simplifying a bit, part of the reason for this is that networking has long been part of the systems curricula in many departments, and thus the Internet, defined via the TCP/IP networking protocols, has long been considered an important part of CS work. The Web, despite having its own protocols, algorithms, and architectural principles, is often viewed by people in the CS field as an application running on top of the Net, more than as an entity unto itself.

This is odd, as the Web is the most used and one of the most transformative applications in the history of computing, even of human communications. It has changed how those in academia teach, communicate, publish, and do research. In industry, it has not only created an entire sector (or, arguably, multiple sectors) but affected the communications and delivery of services across the entire industrial spectrum. In government, it has changed not only the nature of how governments communicate with their citizens but also how these populations communicate and even, in some cases, how they end up choosing their governments in the first place; recall the U.S. presidential debates in which candidates took questions online and through YouTube videos. It is estimated that the size of the human population is on the order of 10^{10} people, whereas the number of separate Web documents is more than 10^{11}.

Computing has made significant contributions to the Web. Our everyday use of the Web depends on fundamental developments in CS that took place long before the Web was invented. Today's search engines are based on, for example, developments in information retrieval with a legacy going back to the 1960s. The innovations of the 1990s [9, 23] provide the crucial algorithms underlying modern search and are fundamental to Web use. New resources (such as Hadoop, lucene.apache.org/hadoop/, an open-source software framework that supports data-intensive distributed applications on large clusters of commodity computers) make it possible for students to explore these algorithms and experiment with large-scale Web-programming practices like MapReduce parallelism [11] in a way not previously accessible beyond a few top universities.

Other aspects of human interaction on the Web have been studied elsewhere. Of special note, many interesting aspects of the use of the Web (such as social networking, tagging, data integration, information retrieval, and Web ontologies) have become part of a new "social computing" area at some of the top information schools. They offer classes in the general properties of networks and interconnected systems in both the policy and political aspects of computing and in the economics

Figure 2: The Web presents new challenges to software engineering and application development.



of computer use. However, in many of these courses, the Web itself is treated as a specific instantiation of more general principals. In other cases, the Web is treated primarily as a dynamic content mechanism that supports the social interactions among multiple browser users. Whether in CS studies or in information-school courses, the Web is often studied exclusively as the delivery vehicle for content, technical or social, rather than as an object of study in its own right.

Here, we present the emerging interdisciplinary field of Web science[5, 6] taking the Web as its primary object of study. We show there is significant interplay among the social interactions enabled by the Web's design, the scalable and open applications development mandated to support them, and the architectural and data requirements of these large-scale applications (see Figure 1). However, the study of the relationships among these levels is often hampered by the disciplinary boundaries that tend to separate the study of the underlying networking from the study of the social applications. We identify some of these relationships and briefly review the status of Web-related research within computing, We primarily focus on identifying emerging and extremely challenging problems researchers (in their role as Web scientists) need to explore.

A large-scale system may have emergent properties not predictable by analyzing micro technical and/or social effects.

What Is It?

Where physical science is commonly regarded as an analytic discipline that aims to find laws that generate or explain observed phenomena, CS is predominantly (though not exclusively) synthetic, in that formalisms and algorithms are created in order to support specific desired behaviors. Web science deliberately seeks to merge these two paradigms. The Web needs to be studied and understood as a phenomenon but also as something to be engineered for future growth and capabilities.

At the micro scale, the Web is an infrastructure of artificial languages and protocols; it is a piece of engineering. However, it is the interaction of human beings creating, linking, and consuming information that generates the Web's behavior as emergent properties at the macro scale. These properties often generate surprising properties that require new analytic methods to be understood. Some are desirable and therefore to be engineered in; others are undesirable and if possible engineered out. We also need to keep in mind that the Web is part of a wider system of human interaction; it has profoundly affected society, with each emerging wave creating new challenges and opportunities in making information more available to wider sectors of the population than ever before.

It may seem that the best way to understand the Web is as a set of protocols that can be studied for their properties, with individual applications analyzed for their algorithmic properties. However, the Web wasn't (and still isn't) built using the specify, design, build, test development cycle CS has traditionally viewed as software engineering best practice.

Figure 2 outlines a new way of looking at Web development. A software application is designed based on an appropriate technology (such as algorithm and design) and with an envisioned "social" construct; it is indeed a contradiction in terms to talk about a Web application built for a single user on a single machine. The system is generally tested in a small group or deployed on a limited basis; the system's "micro" properties are thus tested. In some cases, when more and more people accept the micro system, accelerating "viral" scaling occurs. For example, when Mosaic, the first popular Web browser, was released publicly in 1992, the number of users quickly grew by several orders of magnitude, with more than a million downloads in the first year; for more recent examples, consider photo-sharing on Flickr, video-uploading on YouTube, and social-networking sites like mySpace and Facebook.

The macro system, that is, the use of the micro system by many users interacting with one another in often-unpredicted ways, is far more interesting in and of itself and generally must be analyzed in ways that are different from the micro system. Also, these macro systems engender new challenges that do not occur at the micro scale; for example, the wide deployment of Mosaic led to a need for a way to find relevant material on the growing Web, and thus search became an important applica-

tion, and later an industry, in its own right. In other cases, the large-scale system may have emergent properties that were not predictable by analyzing the micro technical and/or social effects. Dealing with these issues can lead to subsequent generations of technology. For example, the enormous success of search engines has inevitably yielded techniques to game the algorithms (an unexpected result) to improve search rank, leading, in turn, to the development of better search technologies to defeat the gaming.

The essence of our understanding of what succeeds on the Web and how to develop better Web applications is that we must create new ways to understand how to design systems to produce the effect we want. The best we can do today is design and build in the micro, hoping for the best, but how do we know if we've built in the right functionality to ensure the desired macroscale effects? How do we predict other side effects and the emergent properties of the macro? Further, as the success or failure of a particular Web technology may involve aspects of social interaction among users, a topic we return to later, understanding the Web requires more than a simple analysis of technological issues but also of the social dynamic of perhaps millions of users.

Given the breadth of the Web and its inherently multi-user (social) nature, its science is necessarily interdisciplinary, involving at least mathematics, CS, artificial intelligence, sociology, psychology, biology, and economics. We invite computer scientists to expand the discipline by addressing the challenges following from the widespread adoption of the Web and its profound influence on social structures, political systems, commercial organizations, and educational institutions.

Beneath the Web Graph

One way to understand the Web, familiar to many in CS, is as a graph whose nodes are Web pages (defined as static HTML documents) and whose edges are the hypertext links among these nodes. This was named the "Web graph" in [22], which also included the first related analysis. The in-degree of the Web graph was shown in Kleinberg et al.[3] and Kumar et al.[24] to follow a power-law distribution; a similar effect was shown in Broder et al.[10] for the out-branching of vertices in the graph. An important result in Dill et al.[12] showed that large samples of the Web, generated through a variety of methods, all had similar properties—important as the Web graph grows, reported in 2005 to be on the order of seven million new pages a day.[17] Various models have been proposed as to how the Web graph grows and which models best capture its evolution; see Donato et al.[14] for an analysis of a number of these models and their properties.

Along with analyses of this graph and its growth, a number of algorithms have been devised to exploit various properties of the graph. For example, the HITS algorithm[23] and PageRank[9] assume that the insertion of a hyperlink from one page to another can be taken as a sort of endorsement of the "authority" of the page being linked to, an assumption that led to the development of powerful search engines for finding pages on the Web. While modern search engines use a number of heuristics beyond these page-authority calculations, due in part to competitive pressure from those trying to spoof the algorithms and get a higher rank, these Web-graph-based models still form the heart of the critical crawlers and rank-assessment algorithms behind Web search.

The links in this Web graph represent single instantiations of the results of calling the HTTP protocol with a GET request that returns a particular representation (in this case an HTML page) of a document based on a universal resource identifier (URI) that serves as an identifier common across the entire Web. So, for example, the URI http://www.acm.org/publications/cacm typed into a standard Web browser invokes the hypertext transfer protocol (HTTP) and returns an HTML page that contains content describing the publication known as *Communications of the ACM.* Note, however, that the content itself contains other URIs that are themselves pointers to objects that are also displayed (such as icons and images) and that the formatting of the page itself may require retrieving other resources (such as cascaded style sheets) or XML DTD documents. So what we might naively view as a single link from, say, a research group's Web page to an article on a *Communications* page will actually involve a number of requests among a number of servers; at the time of this writing, typing the URI for *Communications* into a browser will cause more than 20 different HTTP-GET requests to occur for seven different types of Web formats. Crawlers can capture these links and create the Web graph as, essentially, a static snapshot of the linking of the Web.

However, the Web graph is just one abstraction of the Web based on one part of the processing and protocols underlying its function. While it is an important result that the Web graph is scale-free, it is the design of the protocols and services that we now call the Web that makes it possible for it to be this way. The Web was built around a set of core design components defined in *The Architecture of The World Wide Web, Volume 1*[21] as "the identification of resources, the representation of resource state, and the protocols that support the interaction between agents and resources in the space."

A feature of the Web is that, depending on the details of a request, different representations may be served up to different requesters. For example, the HTML produced may vary based on conditions hidden from the client (such as which particular machines in a back-end server farm process the request) and by the server's customization of the response. Cookies, representing previous state, may also be used, causing different users to see different content (and thus have different links in the Web graph) based on earlier behavior and visits to the same or to other sites. This sort of user-dependent state is not directly accounted for in current Web-graph models.

There are also other ways the Web, as an application of the Internet, cannot simply be analyzed using the model of a quasi-static graph of linked hypertext pages. For example, many Web sites use Web forms to access a wealth of information behind the servers, where that information, sometimes called "the deep Web," is not visible in the Web model. For many sites, in which the applications's data forms a linked Web, the links are not explicit, and HTTP-POST requests are used instead of the HTTP-GETs in the Web graph. In other cases, these sites generate com-

plex URIs that use GET requests to pass on state[a], thus obscuring the identity of the actual resources.

URIs that carry state are used heavily in Web applications but are, to date, largely unanalyzed. For example, in a June 2007 talk, Udi Manber, Google's VP of engineering, addressed the issue of why Web search is so difficult,[25] explaining that on an average day, 20%–25% of the searches seen by Google have never been submitted before and that each of these searches generates a unique identifier (using server-specific encoding information). So a Web-graph model would represent only the requesting document (whether a user request or a request generated by, for example, a dynamic advertisement content request) linked to the www.google.com node. However if, as is widely reported, Google receives more than 100 million queries per day, and if 20% of them are unique, then more than 20 million links, represented as new URIs that encode the search term(s), should show up in the Web graph every day, or around 200 per second. Do these links follow the same power laws? Do the same growth models explain these behaviors? We simply don't know.

Analyzing the Web solely as a graph also ignores many of its dynamics (especially at short timescales). Many phenomena known to Web users (such as denial-of-service attacks caused by flooding a server and the need to click the same link multiple times before getting a response) cannot be explained by the Web-graph model and often can't be expressed in terms amenable to such graph-based analysis. Representing them at the networking level, ignoring protocols and how they work, also misses key aspects of the Web, as well as a number of behaviors that emerge from the interactions of millions of requests hitting many thousands of servers every second. Web dynamics were analyzed more than a decade ago,[20] but the combination of (i) the exponential growth in the amount of Web content, (ii) the change in the number, power, and diversity of Web servers and applications, and (iii) the increasing number of diverse users from everywhere in the world makes a similar analysis impossible today without creating and validating new models of the Web's dynamics. Such models must also pay special attention to the details of the Web's architecture, as well as to the complexity of the interactions actually taking place there.

Today's interactive applications are very early social machines, limited by the fact that they are largely isolated one from another.

Additionally, modern, sophisticated Web sites provide powerful user-interface functionality by running large script systems within the browser. These applications access the underlying remote data model through Web APIs. This application architecture allows users and entrepreneurs to quickly build many new forms of global systems using the processing power of users' machines and the storage capacity of a mass of conventional Web servers. Like the basic Web, each such system is interesting mainly for its emergent macro-scale properties, of which we have little understanding. Are such systems stable? Are they fair? Do they effectively create a new form of currency? And if they do should it be regulated?

Similarly, many user-generated content sites now store personal information yet have rather simplistic systems to restrict access to a person's "friends." This information is not available to wide-scale analysis. Some other sites must be allowed to access the sites by posing as the user or as a friend; a number of three-party authentication protocols are being deployed to allow this. A complex system is thus being built piece by piece, with no invariants (such as "my employer will never see this picture") assured for the user.

The purpose of this discussion is not to go into the detail of Web protocols or the relative merits of Web-modeling approaches but to stress that they are critical to the current and continued working of the Web. Understanding the protocols and issues is important to understanding the Web as a technical construct and to analyzing and modeling its dynamic nature. Our ability to engineer Web systems with desirable properties at scale requires that we understand these dynamics. This analysis and modeling are thus an important challenge to computer scientists if they are to be able to understand

a. These characters, including ?.#, =, and &, followed by keywords, may follow the last "slash" in the URI, thus making for the long URIs often generated by dynamic content servers.

the growth and behaviors of the future Web, as well as to engineer systems with desired properties in a way that is significantly less hit or miss.

From Power Laws to People

Mathematically based analysis of the Web involves another potential failing. Whereas the structure and use of various Web sites (taken mathematically) may have interesting properties, these properties may not be very useful in explaining the behavior of the sites over time. Consider the following example: Wikipedia (www.wikipedia.org), the online wiki-based encyclopedia, includes more than two million articles in English and more than six million in all languages combined. They are hyperlinked, and it is logical to ask whether the hyperlinks have structure similar to those on the Web in general or whether, since this is a managed corpus, they have yet other properties.

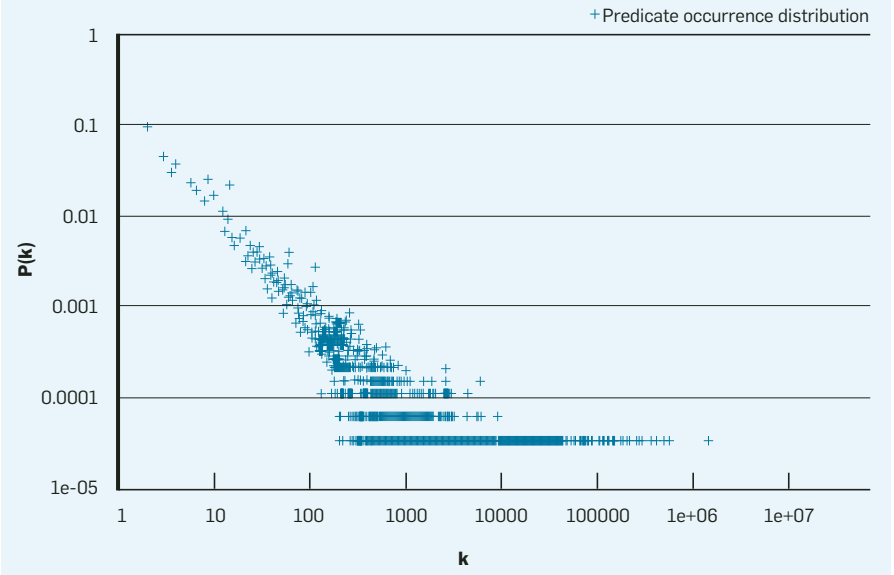
Answering can be done in a number of ways; Figure 3 shows the result of one of them. In this case, DBPedia (dbpedia.org), which is a dump of the link structure of Wikipedia using the labeled links of the resource description framework, or RDF, has been analyzed with respect to the use of the link labels; that is, we are looking at the structure of Wikipedia as opposed to the linguistic content of its pages. The figure shows the same kind of Zipf-like distribution found in the original Web graph analyses. There is also some evidence[16] and a lot of speculation[29] that similar effects can be seen in the use of tags in Web-based tagging systems. Current research is also exploring whether these results depart from such models as preferential attachment[3] used to explain the scale-free features of Web graphs.

Unfortunately, whatever explains these effects, another aspect of Wikipedia's use is not explained by these models and does not necessarily follow from these properties. Wikipedia is built on top of the MediaWiki software package (www.mediawiki.org/wiki/MediaWiki), which is freely available and used in many other Web applications besides Wikipedia. While some of them have also been successful, many have failed to generate significant use. A purely "technological" explanation cannot account for this; rather, something about the organizational structures of Wikipedia and the needs of its users accounts for its success over other systems built from the same code base. The model by which articles are created, edited, and tracked is provided by the underlying technology. The social model enabled by humans interacting in ways allowed by that technology is more difficult to explain. The dynamics of any "social machine" are highly complex, and dozens of academic papers, from multiple disciplines, have been written about it; en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies uses Wikipedia itself to maintain an up-to-date reference list.

The idea of a social machine was introduced in *Weaving the Web*,[8] which hypothesized that the architectural design of the Web would allow developers, and thus end users, to use computer technology to help provide the management function for social systems as they were realized online. The social machine includes the underlying technology (mediaWiki in the case of Wikipedia) but also the rules, policies, and organizational structures used to manage the technology. Examples abound on the Web today. Consider the coupling of the application design of blogging-support systems (such as LiveJournal and WordPress) with the social mechanisms provided by blogrolls, permalinks, and trackbacks that have led to the so-called blogosphere. Similarly, the protocols used by social networking sites like MySpace and Facebook have much in common, but the success or failure of the sites hinges on the rules, policies, and user communities they support. Given that the success or failure of Web technologies often seems to rely on these social features, the ability to engineer successful applications requires a better understanding of the features and functions of the social aspects of the systems.[b]

Today's interactive applications are very early social machines, limited by the fact that they are largely isolated from one another. We hypothesize that (i) there are forms of social machine that will someday be significantly more effective than those we have today; (ii) that different social processes interlink in society and therefore must be interlinked on the Web; and (iii) that they are unlikely to be developed through a single deliberate effort in a single proj-

Figure 3: Results of an analysis of the link structure of Wikipedia with respect to the use of link labels, not the linguistic content of pages.



b. When we say "success" or "failure," we are referring not to the business factors that determine whether, for example, Facebook or MySpace will attract more users but to the success or failure of the sites to provide the particular types of social interaction for which they are designed.

ect or site; rather, technology is needed to allow user communities to construct, share, and adapt social machines so successful models evolve through trial, use, and refinement.

A number of research challenges and questions must be resolved before a new generation of interacting social machines can be created and evolved this way:

- What are the fundamental theoretical properties of social machines, and what kinds of algorithms are needed to create them?;
- What underlying architectural principles are needed to guide the design and efficient engineering of new Web infrastructure components for this social software?;
- How can we extend the current Web infrastructure to provide mechanisms that make the social properties of information-sharing explicit and guarantee that the use of this information conforms to relevant social-policy expectations?; and
- How do cultural differences affect the development and use of social mechanisms on the Web? As the Web is indeed worldwide, the properties desired by one culture may be seen as counterproductive by others. Can Web infrastructure help bridge cultural divides and/or increase cross-cultural understanding?

In addition, a crucial aspect of human interaction with information is our ability to represent and reason over such attributes as trustworthiness, reliability, and tacit expectations about the use of information, as well as about privacy, copyright, and other legal rules. While some of this information is available on the Web today, we lack structures for formally representing and computing over them. Traditional cryptographic security research and well-known access-control-policy frameworks have failed to meet these challenges in today's online environment and are thus insufficient as a foundation for the social machines of the future. Recent work on formal models for privacy[b] has demonstrated that traditional cryptographic approaches to privacy protection can fail in open Web environments. Similar problems with copyright enforcement have also hampered the flow of commercial and scholarly information on the Web.[27] To this end, an exemplar Web science research area we are pursuing involves interdisciplinary research toward augmenting Web architecture with technical and social conventions that increase individual accountability to social and legal rules governing information use.[31] Continued failure to develop scalable models for handling policy will impede the ability of the Web to be the best possible medium for exchanging cultural, scientific, and political information.

The Web is changing at a rate that may be greater than even the most knowledgeable researcher's ability to observe it.

Further, we can see from the dramatic growth of new collaborative styles of creating and publishing information on the Web that many of the social institutions we rely on to judge trustworthiness and veracity are missing from our online information life. Being able to engineer the Web of the future requires not only understanding it as a computational structure but also how it interacts with and supports interaction among its users.

An important aspect of research exploring the influence of the Web on society involves online societies using Web infrastructure to support dynamic human interaction. This work—seen in trout.cpsr.org and other such efforts—explores how the Web can encourage more human engagement in the political sphere. Combining it with the emerging study of the Web and the coevolution of technology and social needs is an important focus of designing the future Web.[30]

The Web of Data

This emerging area of study involves the heavy use of tagging provided by many of what are known as Web 2.0 technologies. Articles, blogs, photos, videos, and all manner of other Web resources may be annotated with user-generated keywords, or tags, that can later be used for searching or browsing these resources. Much has been made of how "folksonomies," or taxonomies that emerge through the use of tags, can be used as metadata to help explain the content of the objects being described.

One aspect of tagging generating interest today is the need for "social context" in tagging.[26] Many tags involve terms that are extremely ambiguous in a general context. For example, first names are popular tags on Flickr,

though they are not good general search terms. On the other hand, in a specific social context (such as a particular person's photos), the same tag can be useful since it can designate a particular individual. The use of a tag as metadata often depends on such a context, and the "network effect" in these cites is thus socially organized.[19]

A more ambitious use of metadata involves recent applications of semantic Web technologies[7] and represents an important paradigm shift that is a significant element of emerging Web technologies. The semantic Web represents a new level of abstraction from the underlying network infrastructure, as the Internet and Web did earlier. The Internet allowed programmers to create programs that could communicate without concern for the network of cables through which the communication had to flow. The Web allows programmers and users to work with a set of interconnected documents without concern for the details of the computers storing and exchanging them.

The semantic Web will allow programmers and users alike to refer to real-world objects—people, chemicals, agreements, stars, whatever—without concern for the underlying documents in which these things, abstract and concrete, are described. While basic semantic Web technologies have been defined and are being deployed more widely, little work has sought to explain the effect of these new capabilities on the connections within the Web of people who use them.[28]

The semantic Web arena reflects two principle nexuses of activity. One tends to involve data (and the Web), and the other on the domain (and semantics). The first, based largely on innovation in data-integration applications, focuses on developing Web applications that employ only limited semantics but provide a powerful mechanism for linking data entities using the URIs that are the basis of the Web. Powered by the RDF, these applications focus largely on querying graph-oriented triple-store databases using the emerging SPARQL language, which helps create Web applications and portals that use REST-based models, integrating data from multiple sources without preexisting schema. The second, based largely on the Web Ontology Language, or OWL, looks to provide models that can be used to represent expressive semantic descriptions of application domains and provide inferencing power for both Web and non-Web applications that need a knowledge base.

Current research is exploring how the databases of the semantic Web relate to traditional database approaches and to scaling semantic Web stores to very large scales.1 In terms of modeling, one goal is to develop tools to speed inference in large knowledge bases (without sacrificing performance), including how to exploit trade-offs between expressivity and reasoning to provide the capabilities needed for Web scale.[15] A market is beginning to emerge for "bottom-up" tools driven by data and "top-down" technologies driven by Web ontologies. Creating back-ends for the semantic Web is being transitioned (bottom-up) from an arcane art into an emerging Web application programming approach, as new open-source technologies integrate well with traditional Web servers. At the same time, new tools support ontology development and deployment (top-down), and tens of thousands of OWL ontologies are available for jumpstarting new domain-modeling efforts. In addition, approaches using rule-based reasoning modified for the Web have also gained attention.[4] Engineering the future Web includes the design and use of these emerging technologies, along with how they differ from traditional approaches to databases, in one case creating back-ends for the semantic Web, in the other new tools for ontology-based applications.

The semantic Web is a key emerging technology on the Web, but, also, as we've discussed, there are different opinions as to what it is best for and, more important, what the macro effects might be. Our lack of a better understanding of how Web systems develop makes it difficult for us to know the kinds of effects the technology will produce at scale. What social consequences might there be from greater public exposure and the sharing of information hidden away in databases? A better understanding of how Web systems move from the micro to the macro scale would provide a better understanding of how they could be developed and what their potential societal effects might be.

Conclusion

The Web is different from most previously studied systems in that it is changing at a rate that may be of the same order as, or perhaps greater than, even the most knowledgeable researcher's ability to observe it. An unavoidable fact is that the future of human society is now inextricably linked to the future of the Web. We therefore have a duty to ensure that future Web development makes the world a better place. Corporations have a responsibility to ensure that the products and services they develop on the Web don't produce side effects that harm society, and governments and regulators have a responsibility to understand and anticipate the consequences of the laws and policies they enact and enforce.

We cannot achieve these aims until we better understand the complex, cross-disciplinary dynamics driving development on the Web—the main aim of Web science. Just as climate-change scientists have had to develop ways to gather and analyze evidence to prove or disprove theories about the effect of human behavior on the Earth's climate, Web scientists need new methodologies for gathering evidence and finding ways to anticipate how human behavior will affect development of a system that is evolving at such an amazing rate. We also must consider what would happen to society if access to the Web was denied to some or all and to raise awareness among major corporations and governments that the consequences of what appear to be relatively small decisions can profoundly affect society in the future by affecting Web development today.

Computing plays a crucial role in the Web science vision, and much of what we know about the Web today is based on our understanding of it in a computational way. However, as we've explored here, significant research must still be done to be able to engineer future successful Web applications. We must understand the Web as a dynamic and changing entity, exploring the emergent behaviors that arise from the "macro"

interactions of people enabled by the Web's technology base. We must therefore understand the "social machines" that may be the critical difference between the success or failure of Web applications and learn to build them in a way that allows interlinking and sharing.

Acknowledgments

Figure 2 is taken from talks Tim Berners-Lee gave in 2007 (www.w3.org/2007/Talks/1018-websci-mit-tbl/Overview.html). We also thank the other members of the WSRI Scientific Council (webscience.org/about/people/) for input relating to the goals of Web science and the interaction of the Web and computer and information sciences. We are indebted to Konstantin Mertsalov of Rensselaer Polytechnic Institute for the DBpedia analysis discussed in the section on power laws. C

References

1. Abadi, D., Marcus, A., Madden, S., and Hollenbach, K. Scalable semantic Web data management using vertical partitioning. In *Proceedings of the 33rd International Conference on Very Large Data Bases* (Vienna, Austria, Sept. 23–27). VLDB Endowment, Heidelberg, 2007.
2. Backstron, L., Dwork, C., and Kleinberg, J. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International World Wide Web Conference* (Banff, Alberta, Canada, May 8–12). ACM Press, New York, 2007.
3. Barabasi, A. and Albert, A. Emergence of scaling in random networks. *Science 286* (1999).
4. Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., and Hendler, J. N3Logic: A logical framework for the World Wide Web. *Theory and Practice of Logic Programming* (2008).
5. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Wietzner, D. Creating a science of the Web. *Science 311* (2006).
6. Berners-Lee, T., Hall, W., Hendler, J., O'Hara, K., Shadbolt, N., and Weitzner, D. A framework for Web science. *Foundations and Trends in Web Science 1*, 1 (Sept. 2006).
7. Berners-Lee, T., Hendler, J., and Lassila, O. The semantic Web. *Scientific American* (May 2001).
8. Berners-Lee, T. and Fischetti, M. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web.* Harper Collins, New York, 1999.
9. Brin, S. and Page, L. The anatomy of large-scale hypertextural Web search engine. Presented at the Sixth International World Wide Web Conference (Santa Clara, CA, Apr. 7–11, 1997).
10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web. In *Proceedings of the Ninth International World Wide Web Conference* (Amsterdam, The Netherlands, May 15–19). Elsevier, Amsterdam, The Netherlands, 2000.
11. Dean, J. and Ghemawat, S. MapReduce: Simplified data processing on large clusters. In *Proceedings of the Sixth Symposium on Operating System Design and Implementation* (San Francisco, Dec. 6–8). USENIX Association, Berkeley, CA, 2004.
12. Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. Self-similarity in the Web. In *Proceedings of the 27th International Conference on Very Large Data Bases* (Rome, Italy, Sept. 11–14). Morgan Kaufmann Publishers, Inc., San Francisco, 2001.
13. Domingos, P., Golbeck, J., Mika, P., and Nowak, A. Social networks and intelligent systems. *IEEE Intelligent Systems, Trends & Controversies 20*, 1 (Jan./Feb. 2005).
14. Donato, D., Laura, L., Leonardi, S., and Millozzi, S. The Web as a graph: How far we are. *ACM Transactions on Internet Technology 7*, 1 (Feb. 2007).
15. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., and Srinivas, K. The Summary Abox: Cutting ontologies down to size. In *Proceedings of the International Semantic Web Conference* (Athens, GA, Nov. 5–9). Springer Berlin, Heidelberg, 2006.
16. Golder, S. and Huberman, B. *The Structure of Collaborative Tagging Systems* (2005); arxiv.org/abs/cs/0508082.
17. Gulli, A. and Signorini, A. The indexable Web is more than 11.5 billion pages. In the special-interest tracks and posters of the 14th International World Wide Web Conference (Chiba, Japan, May 10–14). ACM Press, New York, 2005.
18. Hendler, J. Web 3.0: Semantic Web chicken farms. *IEEE Computer 41*, 1 (Jan. 2008).
19. Hendler, J. and Golbeck, J. Metcalfe's Law, Web 2.0, and the semantic Web. *Journal of Web Semantics 6*, 1 (Feb. 2008).
20. Huberman, B. and Lukose, R. Social dilemmas and Internet congestion. *Science 277*, 5325 (July 1997).
21. Jacobs, I. and Walsh, N. *Architecture of the World Wide Web, Vol. One.* W3C Recommendation, Dec. 15, 2004; www.w3.org/TR/webarch/.
22. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. The Web as a graph: Measurements, models, and methods. In *Proceedings of the Fifth Annual International Conference on Computing and Combinatorics* (Tokyo, July 26–28). Springer, New York, 1999.
23. Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*, 5 (Sept. 1997).
24. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Trawling the Web for emerging cyber communities. In *Proceedings of the Eighth International World Wide Web Conference* (Toronto, May 11–14). Elsevier North-Holland, Inc., New York, 1999.
25. Manber, U. *Why Search Is a Hard Problem.* Presentation at Supernova 2007 (San Francisco, June 16–18, 2008); www.readwriteweb.com/archives/udi_manber_search_is_a_hard_problem.php
26. Marcus, A. and Perez, A. m-YouTube mobile UI: Video selection based on social influence. In *Proceedings of the 12th International HCI Conference* (Beijing, July 22–27). Springer, 2007.
27. Samuelson, P. Copyright's fair use doctrine and digital data. *Commun. ACM 37*, 1 (Jan.1994), 21–27.
28. Shadbolt, N., Hall, W., and Berners-Lee, T. The semantic Web revisited. *IEEE Intelligent Systems 21*, 3 (May/June 2006).
29. Shirky, C. *Power Laws, Weblogs, and Inequality.* In Clay Shirky's blog (2003); www.shirky.com/writings/powerlaw_weblog.html.
30. Shneiderman, B. Web science: A provocative invitation to computer science. *Commun. ACM 50*, 6 (June 2007), 25–27.
31. Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G. Information accountability. *Commun. ACM 51*, 6 (June 2008).
32. Weitzner, D., Hendler, J., Berners-Lee., T., and Connolly, D. Creating a policy-aware Web: Discretionary, rule-based access for the World Wide Web. In *Web and Information Security*, E. Ferrari and B. Thuraisingham, Eds. IRM Press, Hershey, PA, 2006.

Funding for this work comes from the U.S. National Science Foundation (Policy Aware Web and Transparency Aware Data Mining Projects), iARPA(End-to-End Semantic Accountability), the U.K. Engineering and Physical Sciences Research Council (Advanced Knowledge Technologies Project), and the U.S. Army Research Laboratory and U.K. Ministry of Defence (U.S./U.K. Information Technology Alliance). We also thank industrial and individual donors to the authors' research at RPI, Southampton, and MIT and to the Web Science Research Initiative (www.webscience.org).

James Hendler (hendler@cs.rpi.edu) is the Tetherless World Chair of Computer and Cognitive Science at Rensselaer Polytechnic Institute, Troy, NY.

Nigel Shadbolt (nrs@ecs.soton.ac.uk) is professor of artificial intelligence and deputy head of the School of Electronics and Computer Science at Southampton University, Southampton, U.K.

Wendy Hall (wh@ecs.soton.ac.uk) is a professor of computer science at the University of Southampton, Southampton, U.K.

Tim Berners-Lee (timbl@csail.mit.edu) is the Director of the World Wide Web Consortium and holds the 3Com Founders chair and is a senior research scientist in the Laboratory for Computer Science and Artificial Intelligence at the Massachusetts Institute of Technology, Cambridge, MA.

Daniel Weitzner (djweitzner@csail.mit.edu) is director of the Massachusetts Institute of Technology Decentralized Information Group and principle research scientist in the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA.

ILLUSTRATION BY MARIUS WATZ