Web science: a new frontier

Nigel Shadbolt¹, Wendy Hall¹, James A. Hendler² and William H. Dutton³

¹Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
²Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
³Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK

During the past 20 years, humans have built the largest information fabric in history. The World Wide Web has been transformational. People shop, date, trade and communicate with one another using it. Although most people are not formally trained in its use, yet it has assumed a central role in their lives. Scientists and researchers cannot imagine their work without it. Governments interface to their citizens using it. Media are seeing the nature of their industry change because of it. Travel, leisure, health, banking, any sector one can think of are changed by what we have created.

The Web is now ubiquitous, and like all things that become commonplace, we take it for granted. This is true for the great majority of users. Until recently, it was true for researchers too. Over the past few years, there has been a growing recognition that the ecosystem that is the Web needs to be treated as an important and coherent area of study—this is Web science.

It is ‘science’ in the original and broad sense of the term—science as the quest to build an organized body of knowledge. As such, it will need to embrace engineering—the Web is an engineered construct, a set of protocols and formalisms. It will need to embrace the human and social sciences—the Web is a social phenomenon whose vast scale has produced emergent properties and transformative behaviours. The Web is a space built and used by people for people. We will not understand it if we simply reduce it to its technological parts.

In September 2010, we convened a workshop at the Royal Society in London as part of the Society’s 350th anniversary celebrations. We assembled a group of researchers whose work is defining this new area of study. They had backgrounds in mathematics and physics, computer science and engineering, social and human sciences. In this Issue, we publish papers from that workshop. In the period since those talks, the work...
has become even more pivotal to an understanding of the Web. The workshop itself, and hence this Issue, was organized around a number of broad themes. The first of these focused on understanding the form and structure of the Web with talks from Barabási, Kleinberg, Chayes and May.

Emergence is one of the key features of the Web—whether it is the emergence of the ‘blogosphere’ or the appearance of Wikipedia, the increasing linking of scientific data or social networks—complex structures emerge from apparently simple principles. During the past decade, the Web’s connectivity has been studied. We have learnt that wherever you look on the Web some pages have many more links to them than others, and this distribution looks largely the same at whatever scale you sample. Barabási [1] was one of the first to extend and develop the mathematics needed to model the Web—it was he who proposed the scale-free nature of the Web and described some of the power laws it appears to follow. The abstract of his talk contained in this Issue cites those important results.

Barabási [1] has also noted that because the Web is ‘scale-free’ in this way even if a majority of the pages were removed a path from one page to any other is likely to exist. However, removing a relatively small number of the more highly connected items would lead to the disintegration of the network. This has consequences for our understanding of the Web’s resilience and how to secure it from attack. Barabási [1] also discovered that the Web is a ‘small world’—despite the billions of pages you can get from one page to any other in 15 or so clicks. The Web has this beneficial structure of short paths because of human behaviour—we group into communities and to understand this we require insights from sociology and psychology.

It was Kleinberg [2] who produced some of the earliest and most powerful algorithms for finding communities on the Web, and he was also one of the first to characterize its small world nature. In his paper, he discusses some of the observations, theories and conclusions that have grown from the study of Web-scale social interaction. He focuses on issues, including the mechanisms by which people join groups, the ways in which different groups are linked together in social networks. He considers the interplay of positive and negative interactions in these networks. These turn out to be crucial for an understanding of how networks form online.

Chayes [3] uses methods from computer science, mathematics and physics to understand rapidly growing, dynamic graph structures such as the Web. For her ‘all the world’s a graph, and all the people and domains merely vertices’. A graph is represented as a set of vertices V and a set of edges E, so that, for instance, in the World Wide Web, V is the set of pages and E the directed hyperlinks; in a social network, V is the people and E the set of relationships. Chayes [3] outlined in her presentation to the workshop how mathematics can be used to study the Web through four stages: first, modelling online networks as large finite graphs; second, sampling pieces of these graphs; third, understanding and then controlling processes on these graphs; and fourth, developing algorithms for these graphs and using them to improve the online experience. The abstract that summarizes her talk presents the background research that encapsulates this approach to the Web.

A final paper on our understanding of the Web as a network structure comes from the theoretical ecologist Lord May [4]. May’s paper offers a presentation of what is understood about networks in ecological contexts and their dynamical properties. He then describes how these insights can extend to an understanding of systemic risk in communities of networked elements such as banks. This understanding of real-world network phenomenology leads him to issue a number of caveats around assumptions about the structure of the Web. He points out that the Web may not be altogether ‘scale-free’. We cannot assume that a sample from a network is representative of the network’s degree distribution. The challenge here is to know how to sample the Web so as to meet these conditions. He points out that even if the degree distribution is accurately known, it does not fully characterize the network. He also notes that, in many contexts, a network’s dynamical response to, for example, disturbance will depend not only on its topology, but also on the strengths of, or flows along, individual links.

There is clearly much more research needed on the topology and dynamics of the Web if we are to fully understand its form and nature, structure and properties. It is a difficult problem not
least because large amounts of the Web’s content and structure are now created dynamically—they are generated at the point at which a user makes a request of a website. This in turn impacts on trying to understand its structure—do our results apply in this ephemeral Web?

Our second workshop theme, and the papers arising in this Issue, focused on research that looked to the future evolution of the Web as an engineered platform and as a generic computational architecture. It comprised talks from Wu, Robertson, Kermarrec and Karger, all of which are presented in this Issue.

Wu heads the Chinese Next Generation Internet project. The Internet underlies the World Wide Web, but there are numerous challenges in ensuring it remains capable of supporting the Web. These include problems of scalability, guaranteeing high levels of performance, security, real-time adaptability, resilience and mobile communications. The next-generation Internet will have to focus on developing solutions for these challenges. For example, IPv6 provides a large-scale IP address space and makes today’s Internet able to connect not only to computers but also a myriad of electronic devices that will constitute the Internet of Things. IPv6 will be a core protocol and play an important role in the next-generation Internet. In their paper, Wu and co-workers [5] describe the China Next Generation Internet (CNGI) and the technological challenges and opportunities it faces. If we are to maintain a universal architecture and single Web then solutions to the challenges Wu and co-workers describe will be essential.

The paper from Robertson & Giunchiglia [6] describes another global ambition—the aim of ‘programming the global computer’, originally identified by Robin Milner and others as one of the grand challenges of computing research. At the time this phrase was coined, it was natural to assume that this objective might be achieved primarily through extending programming and specification languages. The Internet, however, has brought with it a different style of computation that (although harnessing variants of traditional programming languages) operates in a style different from those with which we are familiar. The ‘computer’ on which we are running these computations is a social computer or a social machine in the sense that humans perform many of the elementary functions of the computations it runs and successful execution of a program often depends on properties of the human society over which the program operates. These sorts of programs are not programmed in a traditional way and may have to be understood in a way that is different from the traditional view of programming. This shift in perspective raises new challenges for the science of the Web and for computing in general—these were discussed in Robertson’s talk and are further expanded in his paper with Giunchiglia [6].

Kermarrec’s [7] paper describes new kinds of distributed architecture for the Web. It is driven by the observation that personalization is a dominant theme on the Web. The Web has become a user-centric platform where users post, share, annotate, comment, forward content be it text, videos, pictures, URLs etc. This social dimension creates tremendous new opportunities for information exchange over the Internet and Web. But how are we to customize the experience so that it is relevant to our particular interests and context. Her paper reviews existing personalization approaches, most of which are centralized. She then advocates the need for fully decentralized systems. These however present two familiar challenges: scalability and privacy. Her paper shows how to achieve personalization in decentralized systems and describes the continuing challenges of providing effective privacy-enhancing technologies.

As the Web grows, changes and expands, researchers must seek novel ways to explore, navigate and search its content. Karger’s [8] paper discusses the challenge of exploring the vast information spaces that are continuously emerging on the Web. The problem is compounded by the fact that, to date, most data authoring and management tools have been oriented towards programmers and Web developers. Users have been unable to really harness structured data for information management and communication. He describes research that enables end users to define their own schemas (without even knowing what a schema is), manage data and author (not program) interactive Web visualizations using the Web tools with which they are already familiar.

A third session at the workshop dealt with our understanding of the Web as a social construct with talks from Levy, Castells, Margetts and von Ahn. It is unfortunate only two of these presentations, Margett’s and von Ahn’s, are represented in this collection of papers. However,
it is worth noting that Levy’s work as a philosopher who researches how humans can exhibit ‘collective intelligence’ has been at the forefront of thought on cyberculture since the 1990s. A summary of his thinking can be found in Levy [9]. Castells is a sociologist and is credited with the first large-scale empirical surveys of the impact of the Internet and Web on Society. He has authored seminal texts on the rise of the information culture [10]. These provide important insights for our understanding of the impact of the Web.

Margetts works as a political scientist who studies the impact of the Web on e-government and seeks to understand the consequences for digital era governance (DEG). As our information culture has evolved so has the way in which governments and citizens interact. Governance in the digital era is an important topic of Web science. The Web has allowed new forms of debate to arise, new ways to engage the citizen or for the citizen to engage with government, new opportunities to furnish the open data of government.

The use of Web-based applications such as social media, online social networking, and wikis, for example, has facilitated peer production, crowd-sourcing, widespread network effects, new organizational forms, and a general "deformalization" of organizations. These developments blur state–societal boundaries. They support a move towards ‘open-book’ governance, transparency and open data initiatives. These hold the promise of co-production and co-creation of government services. Margetts & Dunleavy’s [11] paper in this Issue proposes a DEG model placing Web-based technologies at the centre and replacing previous ‘new public management’ models.

The Web ecosystem is composed of humans and machines—together they are able to solve problems that neither could solve alone. We are trying to understand how to exploit human participation to solve a range of computationally challenging tasks. Tasks that range from classifying astronomical objects (http://www.galaxyzoo.org/) to deciphering faded texts and manuscripts (http://www.google.com/recaptcha/digitizing). Humans use the Web to exhibit ‘collective intelligence’. Collaborative behaviour with light rules of coordination leads to the emergence of large-scale resources such as Wikipedia. The challenge is to technically enable such collective resources and also to understand what drives people to collaborate in such environments.

Capturing the extraordinary power of the crowd is at the heart of von Ahn’s [12] research. Represented in this Issue as an abstract it distils his work on novel ways to engage human participation to solve a range of computationally challenging tasks. His work on ‘games with a purpose’ is about harnessing human time and energy to address problems that computers cannot yet solve. His basic precept is that although computers have advanced dramatically in many respects over the last 50 years, they still do not possess the basic conceptual intelligence or perceptual capabilities that most humans take for granted. By leveraging human skills and abilities in a novel way, he seeks to solve large-scale computational problems and collect training data to teach computers many of the basic human talents. He treats human brains as processors in a distributed system, each performing a small part of a massive computation. Unlike computer processors, however, humans require an incentive in order to become part of a collective computation. One of his key contributions has been to show how to use online games as a means to encourage participation in the process.

The final theme of the workshop focused on the future for the Web with talks from Jain, Contractor, Zittrain and Berners-Lee. As the Web evolves there are those who talk of the ‘Experiential Web’ which is a Web that encompasses and encodes our daily activities and experiences. A Web that is capable of replaying these experiences in a whole range of media. It is a Web in which more and more objects have a Web presence—from refrigerators to buses, items of clothing to scientific sensors. It is this vision of an event-based Web that Jain [13] describes in his paper. He believes that the rise of social networks, their extensive sharing of live status updates (or microblogs) and their use of experiential media (such as photos) will ultimately lead to a Web of events. His paper details the research challenges to be overcome if we are to realize an EventWeb.

Contractor [14] brings a network science view to the social interactions on the Web. His paper examines how the Web has enabled new forms of agile assembly of teams. Studying the
formation of teams on the Web supports a deeper understanding of the motives and dynamics of virtual teams. This helps in accounting for the high levels of innovation and creativity observed in many contexts on the Web. The ambition is to understand how such productivity can be routinely harnessed.

The Web may appear to be a permanent feature of modern life, but there is nothing inevitable about its continued existence. Once again this is a technical as well as a social challenge. There is work that seeks to ensure that the Web remains distributed yet stable, open and at the same time secure. The principles of universal access and non-proprietary, open formats underlie the Web’s basic protocols. Changes to these formats or to the assumption of content accessibility could have far-reaching consequences that need to be understood.

Zittrain [15] in his talk—transcribed here, edited and reproduced—addresses the very pressing questions of ‘will the Web break?’. He surveys the various pressures it is under by reflecting on what makes it currently work and endure. How certain are we that standards and protocols, the arrangements and management of the Web will endure? Certainly, the pervasive sensor and data Web will present its own attendant technical, scientific and social challenges. Once again we are challenged to ask ‘In the Web what should we expect in terms of privacy?’.

As we link increasing amounts of data together—as the Web of Linked Data emerges—new opportunities and new issues will surface. And it was to these issues that the last of our speakers addressed himself. Tim Berners-Lee, the man who gave us the basis of the World Wide Web, reflected in his talk on the emergence of a Web of data. He described the need to link islands of data and to provide a read–write environment with people and machines able to interconnect and interoperate at the data, information and knowledge levels. A summary of his presentation is co-authored with Kieron O’Hara [16].

The Web has had profound social effects. It has empowered groups and individuals in quite novel ways. We should not forget that while 2 billion people enjoy access to the Web—the majority of humanity do not. As we see the Web take hold in the developing world, we are witnessing new and exciting behaviours—new kinds of markets, economic activity and information dissemination.

Understanding the Web will have wide-reaching implications and is on a par with other great scientific challenges such as understanding the climate, our biological nature or the larger Universe. Web science is a new and emergent discipline that is developing its own methods and techniques. But just as with climate science or life science, it must be interdisciplinary drawing insights from mathematics, physics, computer science, psychology, sociology but also law, political science, economics and more. The Web is humanity connected and as such we will need to understand ourselves if we are to understand what we have developed and co-created during the past two decades. If we are to anticipate how the Web will develop we will require insight into our own nature. Web science is not only a new frontier it is an endeavour that will bring together a new generation of enquiring minds.

We owe particular thanks to those at the Royal Society who helped organize the original workshop and have worked on producing this Issue. Thanks also to Susan Davies for helping ensure both a successful workshop and providing administrative support. Finally, to Kieron O’Hara for providing a series of excellent summaries of work for inclusion in this Issue.

References