1

Introduction

This chapter aims to provide a context for web social science, introducing some of the major themes that are addressed elsewhere in the book.

Section 1.1 provides an introduction to the key technologies and governance structures that underlie the Internet and the web, and presents a timeline of key events (from the perspective of web social science) in the history of the web. Section 1.2 introduces examples of online computer-mediated interaction which feature throughout the book. Section 1.3 introduces three important phases in the conceptualisation of the web: cyberspace, virtual communities and online social networks. Section 1.4 outlines four disciplinary approaches for conducting empirical research using data from the web. Section 1.5 introduces the concept of construct validity in the context of web data. Finally, Section 1.6 looks at whether the web can should be viewed as a tool that people use for achieving social, political and economic outcomes, rather than a force that shapes behaviour.

1.1 THE WEB: TECHNOLOGY, HISTORY AND GOVERNANCE

The starting point for a book on web social science is necessarily a brief introduction to the technology that underlies the web. While the average social scientist will not need to know much about the technology of the web, it is important to know, for example, that the web and the Internet are not synonymous. The Internet came before the web, and the web is in fact built on top of the Internet.

The Internet is a massive, distributed network of computers, originally developed in the US in the 1960s with funding from the Defense Advanced Research Projects Agency (DARPA). Data that is transferred between computers on the Internet is split into relatively small blocks ('packets') which are then reconstituted at the final destination. Packets follow the most efficient pathway to the final destination; if a particular computer is not available they are automatically rerouted. This enables efficient transfer of data

and also means that packets can be delivered even if parts of the network are not functioning (the original interest of DARPA was in ensuring communications in the event of war).

For the packets to be successfully sent and received there need to be rules or *protocols* – two critical protocols are the Transmission Control Protocol (TCP) and the Internet Protocol (IP), jointly referred to as TCP/IP.[1] But TCP/IP are not the only important protocols. The delivery of email involves an additional protocol called the Simple Mail Transfer Protocol (SMTP). The *World Wide Web* (or *web*) is a massive distributed network of *resources* – documents, sounds, images (Box 1.1). The protocol that underlies the web is the HyperText Transfer Protocol (HTTP), which allows the development of web pages written in the HyperText Markup Language (HTML) coding language; these are used to access information on the web. The web is therefore built on top of the Internet. While the Internet is a network of computers connected by cables, the web is a network of documents connected by hypertext links.

The word *network* is very important – a major aim of this book is to show that web social science is network-based social science. However, the networks that are discussed in the book are not networks of computers or documents, but networks of individuals, groups and organisations. That is, the web allows individuals, groups and organisations to form and maintain networks and, in doing so, create digital trace data that can be studied by social scientists. While it is relatively easy to conceive of Facebook as a network of individuals, this book shows that other web applications also facilitate networked behaviour.

The web, which is regarded by some as being the 'largest human information construct in history',[2] was invented by Tim Berners-Lee while based at CERN and was publicly released in 1991. Box 1.2 presents a list of important milestones in the development of the web. The focus is on events that are important for web social science, and references to relevant chapters and sections in the book are provided.

The web is commonly understood to have had three overlapping phases of development or eras: Web 1.0, Web 2.0 and Web 3.0 (Box 1.3). Under Web 1.0, webmasters create content that is then read or consumed by users. Web 1.0 websites are sometimes referred to as comprising the *Static Web* since they typically do not allow a lot of interactivity and the information

[1]An *internet* (lower-case 'i') is any set of computer networks connected by TCP/IP. The *Internet* (upper-case 'i') is the largest set of networks – this is the open and public set of computer networks that we all use. An internet within a single organisation is called an *intranet* (although, technically, this refers to a set of networks that could be using any protocol, not necessarily TCP/IP).

[2]http://webscience.org/webscience.html

presented (often reflecting organisational goals, products, services) does not change regularly (relative to the constant flow of change on sites such as Facebook and Twitter).

Web 2.0 blurs the distinction between webmasters and users, with blogging tools, social network sites (e.g. Facebook) and microblog services (e.g. Twitter) enabling non-technical people to both produce and consume content. The act of a person both consuming and producing web content has been referred to as 'prosumption' (e.g. Ritzer and Jurgenson, 2010) and 'produsage' (e.g. Bruns, 2008).

BOX 1.1 RESOURCES ON THE WEB

So how are resources on the web found? Resources such as websites are identified via unique numeric IP addresses that consist of four numbers (between 0 and 255) separated by dots. The Domain Name System (DNS) translates an easier-to-remember, character-based, *fully qualified domain name* (also known as the *hostname*, *sitename* or *subdomain*), which is the unique name by which a computer is known on a network, into an IP address.

The hostname comprises two parts (joined by a '.'): the name of the *host* (this is the computer that is connected to the network) and the *domain name*. A domain name usually consists of two parts. A *top-level domain* (TLD) identifies the type of organisation. There are two types of TLD: generic TLDs (e.g. '.com', '.edu') and country-code TLDs (e.g. '.au', '.uk'). A *second-level domain* such as 'google' or 'yahoo' identifies the organisation.

For example, the hostname voson.anu.edu.au consists of the host 'voson' and the domain name 'anu.edu.au', and currently translates (via DNS) into the IP address 150.203.224.58. The generic TLD is '.edu', the country-code TLD is '.au', and the second-level domain is 'anu'.

A *uniform resource locator* (URL) is an address that defines a route to a file on an Internet server (e.g. web server, FTP server). The first part of the address is the protocol identifier, while the second part is the resource name, with the first and second parts being separated by '://'. Thus, the URL http://voson.anu.edu.au/index.html consists of the protocol identifier 'http' indicating that this is a resource that is hosted on a web server, and thus requires HTTP to access it, and the resource name is 'voson.anu.edu.au/index.html'. The resource name is composed of the hostname ('voson.anu.edu.au'), the directory path to the file ('/'), and the file ('index.html').

A *subsite* is a collection of pages within a particular website. For example, the subsite http://voson.anu.edu.au/news is a part of the VOSON project website and contains pages with details on project activities, e.g. http://voson.anu.edu.au/news/2012, http://voson.anu.edu.au/news/2011.

BOX 1.2 WEB TIMELINE

1983 – TCP/IP implemented
1984 – William Gibson publishes *Neuromancer* (Section 1.3.1)
1985 – Domain Name System (DNS) introduced
1990 – The Internet comprises over 100,000 hosts
1991 – Linus Torvalds begins work on the open source Linux operating system (based on the MINIX variant of the Unix operating system) (Section 9.1.1)
1990–1994 – New content-publishing services released, e.g. news/bulletin boards, FTP, gopher (menu-driven system for accessing files), first content search engines (e.g. Brewster Kahle's Wide Area Information Service, WAIS)
1991 – Tim Berners-Lee's World Wide Web is publicly released. The web eventually swamped all other content publishing services
1994 – Netscape web browser released
1997 – Internet Archive starts archiving the web, currently available via the Wayback Machine (Section 4.3.2)
1998 – Sergey Brin and Larry Page publish (and patent) their 'PageRank' search algorithm, paving the way for Google (Section 7.1.1)
Mid-2003 – There are an estimated 180 million registered hosts on the Internet, 40 million websites and between 600 and 700 million users
2003 – Linden Labs launch Second Life virtual world (Section 9.3.2)
2004 – Mark Zuckerberg founds Facebook.com, heralding the rise of social network sites (Sections 3.3.3, 5.1.2)
2004 – Political bloggers play prominent role in US Presidential election (Section 7.3)
2005 – YouTube video-sharing website launched
2006 – Twitter microblogging service launched (Section 5.2.2)
2007 – iPhone launched by Apple, igniting the market for smartphones (Section 1.3.1)
2011 – Social media play prominent role in the Arab Spring and the Occupy Movement (Section 8.2)

BOX 1.3 PHASES IN THE EVOLUTION OF THE WEB

Web 1.0: *Static Web*. Key languages/protocols: HTML, HTTP. Key applications: websites (hosted by web server software such as Apache), web browsers (e.g. Firefox).
Web 2.0: *Collaborative Web*. Key languages/protocols: AJAX, RSS, SOAP, XML. Key applications: web blogs, social network services, microblogs, smartphone operating systems (e.g. Android), software as a service (e.g. Google Docs).
Web 3.0: *Semantic Web*. Key languages/protocols: RDF, SWRL, SPARQL. Key applications: semantic databases, intelligent personal agents.

Web 3.0, or the Semantic Web, involves technologies that make the web more machine-readable, leading to a 'web of data', which is an evolution of the Web 1.0 'web of documents' (Shadbolt et al., 2006). While the technologies underlying the Semantic Web are proven, there is yet to be a general take-up of Web 3.0. While it is possible to retrofit existing websites to make them Web 3.0 compatible, this would entail a massive amount of work, so webmasters are unlikely to do this until there are clear benefits or reasons to do so. The exception is the government sector, where Open Data initiatives are drawing on Web 3.0 technologies. But for the vast majority of the web, while Web 1.0 and Web 2.0 are ubiquitous, Web 3.0 is still in its infancy.

A common feature of all three phases of the web is the use of technologies to help people find the content they want. With Web 1.0, and to a lesser extent Web 2.0, the core enabling technology are the hyperlink, which enables users to efficiently move around the web ('web surfing'), and search engines that index web content and present search results to users. In contrast, Web 3.0 envisages intelligent personal agents finding content on behalf of users by drawing on users' preferences and browsing habits.

Governance of the Internet occurs at two levels: architecture and operation.[3] In relation to architecture, design and refinement of protocol specifications is undertaken by various working groups coordinated by the Internet Engineering Task Force (IETF). Other organisations take specific roles in particular areas. For example, issues relating to transmission media are handled by the Institute of Electrical and Electronic Engineers (IEEE) and the International Telecommunications Union (ITU), and protocols to do with the web are the province of the World Wide Web Consortium (W3C) industry association. The main organisation involved with Internet operation governance is the Internet Corporation for Assigned Names and Numbers (ICANN), which coordinates the DNS, IP addresses and the generic and country code TLD system.

1.2 EXAMPLES OF ONLINE COMPUTER-MEDIATED INTERACTION

This section aims to familiarise readers with several forms of online computer-mediated interaction. The list is not complete, with a focus on the types of online interaction that are discussed elsewhere in this book.

[3]While Internet governance is not a focus of this book, governance structures are looked at in Section 9.1.1 in the context of peer production. Also see the discussion on Internet censorship in Section 8.2.2.

Threaded conversations: newsgroups, discussion groups and chat rooms

Newsgroups are repositories of emails set up for different topics, often hosted on the Usenet system (an example is rec.pets.cats – a Usenet newsgroup dedicated to discussing pet cats). *Threaded conversations* occur within newsgroups when individuals make posts to newsgroups (thus starting a 'thread'), and respond to the posts of other people. Discussion groups (or chat rooms) are hosted on the web and are often functionally similar to newsgroups (which do not necessarily involve web technologies). They can be moderated or unmoderated. An example is the chat rooms that are hosted on America Online (AOL). Another example is Slashdot – a popular web-based technology-related forum, with articles and comments from readers. Slashdot has developed its own subculture involving the accumulation of 'karma' scores, with volunteer moderators being selected from those with high scores. Threaded conversations are looked at in Sections 3.3.2 and 9.1.2.

Web 1.0 websites

A static website is the 'face' of Web 1.0. These generally represent organisational web presence, rather than the web presence of an individual person, and they often do not allow for readers of the website to interact with the website authors.[4] Web 1.0 websites are looked at in Chapters 4 and 6, and in Sections 7.1, 8.1 and 9.2.

Blogs

A *weblog*, or *blog*, is a chronologically updated website, typically written by a single author and designed to provide regular commentary on particular topics or else to serve as an online diary. Technically, there is no difference between a static website and a blog: the differences are in how the site is used. However, an innovation that was developed in the context of blogs is RSS feeds which allow blog subscribers to know when new content has been posted. Blogs are looked at in Section 7.3.

Wikis

A wiki is a website where web pages can be edited by members of the public, using a simplified markup language. Wikis are designed to enable non-technical people to jointly collaborate on the creation of web content, with the best-known example of a wiki being Wikipedia. Wikis are looked at in Section 9.1.2.

[4]Note that organisational websites are increasingly incorporating Web 2.0 features (e.g. blogs and RSS feeds), so the boundaries between Web 1.0 and Web 2.0 are blurring.

Social network sites

Social network sites are websites that allow people to create personal profiles and interact with other people with profiles by requesting and accepting 'friendships' and joint membership in groups (representing people who graduated from a particular university, for example, or who share interests). The best-known example of a social network site is Facebook, but notable predecessors were Friendster (in the US) and Cyworld (in Korea). Other examples of social network sites are LinkedIn (for professional networking) and Renren in China. Social network sites are looked at in Sections 3.3.3 and 5.1.2.

Microblog sites

A microblog allows subscribers to broadcast short messages (e.g. a maximum of 140 characters) to other subscribers of the service. The best-known microblog is Twitter, and Sina Weibo is a prominent example of a Chinese microblog ("weibo"). Microblogs are looked at in Sections 3.3.4 and 5.2.2.

Virtual worlds

Virtual worlds are simulated environments where individuals can assume digital representations (avatars) and interact with other individuals. There are two types of virtual worlds. Massive multiplayer online role-playing games (MMORPGs) are typically fantasy-themed and are derived from earlier 'pencil and paper' role-playing games such as Dungeons and Dragons and the first (entirely text-based) virtual worlds, multi-user dungeons. Individuals assume characters (e.g. human, elf), go on quests with other players (e.g. fight monsters and get treasure), use treasure to buy equipment (e.g. armour, weapons) and gain 'experience points' giving the character greater skills. Examples are EverQuest (EQ), published by Sony Online Entertainment, and World of Warcraft (WoW), published by Blizzard Entertainment. The second type of virtual world is exemplified by Linden Lab's Second Life, which is a popular non-gaming virtual world. In Second Life people can build alternative realities online and, while there are rules governing how you construct your avatar and buildings and how you interact with other people, unlike MMORPGs, Second Life inhabitants are not playing a game. Virtual worlds are looked at in Sections 8.3 and 9.3.2.

1.3 CYBERSPACE, VIRTUAL COMMUNITIES AND ONLINE SOCIAL NETWORKS

This book aims to show how empirical social science can provide insights into the impact of the web on society, and how web data can be used to answer long-standing social science research questions. But the Internet is just infrastructure, and the protocols and services that underpin the web are just tools.

In order to achieve the objectives of the book, we need to go beyond the technology and look at how the web is being used by people and organisations, what type of behaviour is occurring on the web, and how this might reflect real-world behaviour and potentially have real-world impacts.

A starting point is a review of three important phases in the conceptualisation of the web: cyberspace, virtual communities and online social networks. As with the technological phases of the web outlined in Box 1.3, the conceptual phases of the web are overlapping.

1.3.1 Cyberspace

The term *cyberspace* was conceived by the science fiction author William Gibson:

> Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts ... A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding. (William Gibson, *Neuromancer*, 1984, p. 69)

Cyberspace has become a *de facto* synonym for the Internet, and has been hugely influential – especially with academics and activists – as a way of describing the Internet as a virtual place in which people interact[5]. Thus, websites are metaphorically said to exist 'in cyberspace' and any interactions between people would similarly be occurring in cyberspace, rather than in the countries where the participants or website servers are located.

Mapping cyberspace

William Gibson's original concept of cyberspace was a visual one, and there is a huge body of work focused on ways of visualising or mapping cyberspace – see Dodge and Kitchin (2000) for an early compilation. Many of the earlier attempts at mapping cyberspace, while technically impressive and visually striking, did not provide much insight for social scientists since they were large-scale maps of the Internet infrastructure.

With the advent of Web 2.0, it has become increasingly common to see academic research featuring maps of the connections between people (e.g. bloggers, Twitter users, Facebook users) and such maps have the potential to be powerful and evocative displays of social processes. For example, the Divided They Blog image (Adamic and Glance, 2005) which is discussed further in Section 7.3 is a powerful visualisation of political homophily.

However, such visualisations are really just a first step in empirical social science research. They are useful for capturing attention and explaining data

[5]http://en.wikipedia.org/wiki/cyberspace

structure, but need to be supplemented with quantitative empirical network techniques such as exponential random graph models (Section 3.2.2). Also, there has perhaps been too much focus on visualisation at the expense of development of appropriate theoretical frameworks. As noted by Janetsko (2009, p. 170), in many cases 'work centering around nonreactive [online] techniques more or less exclusively addresses visualization of phenomena that are perhaps not properly understood'.

The cyberspace ethos

A *cyberspace ethos* developed during the pioneering years of the Internet, and has been influential in forming attitudes and behaviour in relation to interactions on the Internet. An aim of this book is to provide a framework for understanding whether this ethos exists today and how it complements or conflicts with other norms, laws and institutions for the real world.

Clarke (2004) identified several aspects of the cyberspace ethos:

- *Interpersonal communications.* Cyberspace is viewed as being for interpersonal interactions, with organisations having the roles as providers of resources or services rather than participants. Interpersonal communications on the web were greatly enhanced with the advent of Web 2.0, and consequently a lot of web social science is about individual behaviour on the web. However, social scientific web research has also focused on organisational behaviour on the web (see Chapter 6).
- *Internationalism and universalism.* Although the Internet was developed in the US, content and connectivity are technically available to anyone – there are no borders in cyberspace. However, social science research has focused on the digital divide which was traditionally about borders or boundaries preventing equal access to the web (e.g. DiMaggio et al., 2001). But even if everyone has equal access to the web (and equal skills, so no 'hidden digital divide'), while web content might be equally retrievable, it is not equally visible (because of the role of search engines) and this is looked at in Section 7.1.
- *Egalitarianism.* While participants might have particular roles such as moderators on lists, there is no hierarchy of authority on the Internet, and people behave as though they are, by and large, equal. But authority and hierarchy do play out on the Internet (e.g. O'Neil, 2009). Also the network structure of the web does mean that actors have unequal network positions and hence may experience different behaviour and outcomes. This is looked at later in the context of online collective identity (Section 6.3), reconfiguring access to academic information (Section 9.2.2), and structural holes in Second Life (Section 9.3.2).
- *Openness.* The Internet's fundamental protocols and standards are open to anyone. However, authors such as Zittrain (2008) have argued that the openness of the Internet is under threat with proprietary services such as Apple's iPhone and 'walled gardens' that prevent people moving across social network sites.

- *Communitarianism and mutual service.* Many participants feel that they belong to a community – they both contribute to this community and draw from it. To what extent is the concept of community useful for understanding behaviour on the Internet (Section 1.3.2)? Economic research into open source communities suggest that, rather than altruism, participants may expect deferred benefits (e.g. labour market reputation) – see Section 9.1.2.
- *Freedoms.* A core aspect of the cyberspace ethos is the importance of personal freedom in cyberspace. In fact, many believe that there should be greater personal freedom in cyberspace than in the real world, and they resent activities by governments and corporations to constrain freedom (e.g. censorship and copyright). This has been famously captured in John Perry Barlow's 1996 *A Declaration of the Independence of Cyberspace*:

  > Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. ... I declare the global social space we are building to be naturally independent of the tyrannies you seek to impose on us. You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear.[6]

- But to what extent is it reasonable to expect to have more freedom on the Internet than in the 'real world'? How does this aspect of the cyberspace ethos come into conflict with a core function of government: authority (Section 8.2)?

1.3.2 Virtual communities

While cyberspace is an evocative and influential concept, it does not help us to establish a framework for quantitative analysis of online behaviour. From a research perspective, we are interested in being able to operationalise the concept of cyberspace as a virtual place in which people interact. The most common term (other than cyberspace) used to describe this virtual place where people interact is *virtual community*, which was introduced by Rheingold (1993) and is now seen as being analogous to *online community*. In this section, we look at the concept of virtual community and see how it relates to the concept of community as developed in sociology.

Membership of *social groups* or categories can be defined on the basis of personal or individual characteristics such as ethnicity or sex. Social groups are therefore objectively defined: you are either in or out of the group, and group membership does not necessarily involve interpersonal relations.

[6] https://projects.eff.org/~barlow/Declaration-Final.html

When can we call a group of people a *community*? For community to exist there needs to be a sense of group attachment and belonging – a shared sense of 'one-ness' or 'we-ness' that can be referred to as *collective identity*. So one definition of community is that it is a group of actors who share a collective identity. But the definition is circular, because at the same time a community can be viewed as a vehicle for the emergence of collective identity, via providing common beliefs, norms and shared understandings (Durkheim, 1964).

How does a group of people develop common beliefs, norms and shared understandings? Three important factors have been identified. First, there is a high degree of perceived homogeneity among members on important criteria such as ethnicity and religion (Gusfield, 1975). Second, there is physical proximity or co-location of individuals (e.g. in villages or neighbourhoods). Third, there is the existence of social relations or ties between actors. Taylor (1982) argues that there need to be direct, multiplex and durable relations that are governed by reciprocity and strong interdependence between members. Barry Wellman defines a community as 'networks of interpersonal ties that provide sociability, support, information, a sense of belonging, and social identity' (Wellman, 2001, p. 228). Wellman's view on community is particularly relevant here, with his emphasis on networks of interpersonal ties and no mention of physical co-location of actors as a pre-condition for community. As put by Wellman, 'I do not limit my thinking about community to neighbourhoods and villages' (p. 228).

It is important to emphasise that sociologists have traditionally regarded shared interests as not being sufficient for the existence of community. Durkheim (1964) argued that shared interests are not enough, and there need to be ties based on emotions, while Weber (1922) emphasised the need for feelings of group attachment. However, there is the more recent concept of *community of interest*, where all that the members share is a common interest – they do not necessarily exhibit emotional attachment to or form social ties with other members of the community.

What is the definition of an *online group*, and when can it be called an *online community* or *virtual community*? How does the above definition of community translate into the online world?

An *online group* can be defined as a group of people who conduct personal computer-mediated interactions, where interaction is focused on a topic that reflects the common interests of the group. Drawing on the above discussion on community, an online group is therefore a group of people with shared interests who communicate via the Internet, but where collective identity does not exist. It should be noted that others use the term online or virtual community for what we are calling here 'online group' (the above definition in fact draws from the definition of online community used by Matzat (2004b)). However, we use the term 'online group' here as it more clearly indicates the absence of collective identity. An

online group can therefore equivalently be referred to as a *virtual community of interest*.

A *virtual community* is an online group where there are additionally shared values, norms and understandings. What is the role of the three factors identified above as being important for the formation of collective identity (and hence community) – homogeneity, proximity and social ties – in the formation of virtual community? The role of homogeneity (on the basis of characteristics such as race, religion and ethnicity) is surely diminished since people can interact online without revealing much about these personal characteristics. The importance of physical proximity has also been greatly reduced: the Internet allows people located anywhere in the world to connect.

That leaves us with social ties, and at first glance one might argue that this factor alone has retained its importance in the formation of online communities, and has possibly even gained importance (given the potential diminishing of the roles of homogeneity and proximity). Barry Wellman famously declared that 'a computer network is a social network' (Wellman, 2001, p. 227). Rheingold (1993, p. 5) defines an online community as a group of people who hold computer-mediated discussions on a topic for a sufficiently long time with sufficient emotional involvement, and who form relationships: 'Virtual communities are social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace.'

But it should be finally noted that a virtual community might exist even in the absence of homogeneity, proximity and social ties if the topic of interest that draws the community together is one that itself is value-driven, that is, the members would not be interested in the topic if they did not share common values. This is best explained using an example. The rec.pets.cats newsgroup (where people discuss their cats, i.e. how to care for them) is a good example of an online group while alt.non.racism (a newsgroup devoted to discussing racism, presumably from the point of view that it is morally wrong and should not be present in modern society) is an example of a virtual community.

1.3.3 Online social networks

With the rise of Facebook and other social media, the term 'online social networks' has become increasingly popular. In his revised book on the virtual community, Rheingold (2000) states that had he read work by Barry Wellman earlier, he would have used the term 'online social network' instead of 'virtual community'.

The formal definition of an online social network is covered in Chapter 3, but we note here that a distinction can be made between the terms 'social network site' and 'online social network'. The former refers to an online

environment such as Facebook, while the latter refers to the formal representation of a social network, where the data on ties and nodes are the result of online interactions between individuals (perhaps in a social network site such as Facebook).

Similarly, it is important to emphasise that the terms 'virtual community' and 'online social network' are not synonymous. On the one hand, it is possible to conceive that every virtual community can be represented as an online social network. As Wellman (2001) put it: 'Although not every network is a community – unless you think of NATO or interlocking corporate structures as communities – every interpersonal community is a network.' Thus, using the definition of 'virtual community' above, we would define a newsgroup focused on preventing racism as a virtual community and it would be possible to represent this as an online social network since we could collect the data to represent it as a threaded conversation network (Section 3.3.2).

But it is not the case that an online social network will necessarily be a virtual community. For example, if we extracted a network of real-world friends from their Facebook profiles, we could represent and analyse the data as an online social network (Section 3.3.3). But this would not be an example of a virtual community since it is not the case that these people necessarily share common values and norms leading to collective identity.

Finally, as discussed further in Chapter 3, the term *online social network*, while less ambiguous than 'virtual community', is still not without its difficulties in terms of definitions. In particular, we need to distinguish between the types of connections that exist between participants and whether these are likely to lead to the interdependencies between people that are the hallmark of social networks.

1.4 DISCIPLINARY APPROACHES TO RESEARCHING THE WEB

This section outlines four major disciplinary approaches to conducting empirical research into the web: network science (as practised by applied physicists and computer scientists), network science (as practised by social scientists), information science and media studies. The aim is to give a brief introduction to the various approaches, and then indicate where they are covered in more detail elsewhere in the book. It should be emphasised that we focus here on what sets the disciplinary approaches apart rather than identifying what they have in common. But it should be noted that the boundaries between these approaches are not 'hard' (there is active cross-over), and they might in fact be contested by people working in these areas.

Network science (applied physics and computer science)

We distinguish two variants of network science. The first is that practised by applied physicists and computer scientists who study large-scale networks, with the aim of: (1) measuring the properties of these networks (generating 'stylised facts' or 'empirical regularities') and (2) using statistical-mechanical models to generate simulated networks exhibiting the properties that are observed in real networks.

For example, Barabási and Albert (1999) observed the existence of *power laws* in large-scale networks – many network participants have few or no connections, while a handful are very connected. They explained the emergence of power laws using the concept of *preferential attachment*: in a growing network, new entrants to the network prefer to connect with network participants that are already well connected, thus leading to a 'rich-get-richer' phenomenon. See Section 7.1.1 for further details.

Information science

Webometrics (also known as 'cybermetrics') is an approach for analysing hyperlink data and website usage patterns, drawing on bibliometrics and informetrics (which are subfields of information science). See, for example, Almind and Ingwersen (1997), Björneborn and Ingwersen (2004) and Thelwall et al. (2005).

As discussed in Section 9.2.1, webometrics often involves the use of statistical techniques in an attempt to identify what characteristics of a website and of the people who run the website lead to the acquisition of hyperlinks. In a recent example of webometric research, Barjak and Thelwall (2008) analysed counts of inbound hyperlinks to the websites of life science research teams in order to assess the role of hyperlinks as science and technology output indicators.

Media studies

Media studies is an academic field that draws on both social science and humanities, and is concerned with media content and impact (with a particular focus on mass media). While the term 'network' may be used in a metaphorical way, the media-studies perspective on the web is often characterised by an absence of formal network techniques. Researchers from media studies are often more focused (compared with other social scientists) on the Internet as a transformative technology, that is, the creation of 'citizen journalists' (e.g. Flew, 2007; Goode, 2009) who are challenging old media and transforming (or in some cases, creating) democracy across the globe.

While media-studies research into the web typically does not use formal network techniques, the concept of *issue networks* (e.g. Rogers, 2010a) has been developed as a way of understanding how individuals and organisations

are using the web to engage with particular issues. However, Rogers (2010b, p. 8) points out that 'issue networks may be distinguished from popular understandings of networks, and social networking, in that the individuals or organizations in the network neither need be on the same side of an issue, nor be acquainted with each other (or desire acquaintance)'. See Section 4.2.2 for further details.

Network science (social science)

The second variant of network science is the one practised by social scientists. Note that social science is being defined here as including sociology, political science and economics. This is a narrow definition as many consider media studies to be a social science and some information scientists regard themselves as social scientists.

How does the social science approach to studying the web differ from the other disciplines? First, compared with applied physicists and computer scientists, social scientists are more concerned about using models of behaviour that are clearly grounded in social science. While the preferential attachment model of Barabási and Albert (1999) generates networks that exhibit the power laws that have been found in large-scale networks such as the web, it is a statistical-mechanical model and the actors or agents in the model do not exhibit behaviour that is realistic from the perspective of a social scientist.

Second, compared with researchers from media studies, social scientists are more focused on how the Internet is used by actors (people, organisations, governments) to pursue social, economic and political ends rather than the Internet as a force that is controlling or changing people's behaviour.

Finally, researchers from information science use webometric techniques which allow the finding of answers to the question 'What are the qualities of the actors receiving the most hyperlinks?', while a more social scientific approach to studying hyperlinking behaviour involves the use of exponential random graph models which can answer the question 'Why do actors make or receive a hyperlink?' (Section 4.2.3).

Areas of cross-over

There are of course examples where there is disciplinary cross-over in web research. For example, in the context of studying the visibility of various political messages on the web, Hindman et al. (2003) identified the existence of power laws in the distribution of inbound hyperlinks to web pages containing political content. This is therefore an example of applied physics being used in political science (Section 7.1.1). Similarly, Escher et al. (2006) used techniques from webometrics and network science to study how the web has changed government nodality (the property of being at the centre of social and information networks) – see Section 8.1.

1.5 CONSTRUCT VALIDITY OF WEB DATA

Many concepts in social science are subjective, and it is sometimes difficult to know whether a variable is adequately correlated with the phenomenon of interest that it purports to measure. For example, while IQ ('intelligence quotient') test scores are widely used as proxies for intelligence, some people challenge the *construct validity* of the IQ test: does it measure what is intended (intelligence) or are other factors (education, socio-economic status, culture) going to influence the score to the extent that it diminishes its use as a measure of intelligence?

The construct validity of web data (in particular, digital trace data) is integral to web social science. If one is not able to either empirically or theoretically demonstrate the construct validity of web data for social science research, then one is left wondering why one should, as a social scientist, care about hyperlinks, tweets, Facebook friendships, etc.

This book shows how the construct validity of web data can be assessed in three ways. First, the construct validity of web data may be assessed by testing whether the online network displays *structural signatures* that are consistent with those displayed by real-world actors. For example: Does Facebook friendship network data display homophily on the basis of race, ethnicity (Section 5.1.2)? Are divisions between different groups in the environmental social movement evident in hyperlink networks (Section 6.3)? And to what extent is political affiliation reflected in political blog networks (Section 7.3)?

Second, it may be possible to assess construct validity by testing whether variables constructed from web data are correlated with other accepted measures of the construct. For example, if counts of inbound hyperlinks to academic project websites are correlated with other characteristics of academic teams (e.g. publications, industry connections) that are used as proxies of academic authority or performance, then this is evidence of the construct validity of hyperlink data in the context of scientometrics (Section 9.2.1). In Section 7.1.1 the construct validity of hyperlink data is assessed in the context of the visibility of political information. The argument is made that counts of inbound hyperlinks are likely to be correlated with numbers of visitors to websites ('eyeballs'), and to the extent that the latter is an accepted measure of political visibility, the former therefore has construct validity.

Finally, the construct validity of web data may be demonstrated if it can be shown that an actor's position in an online network has influence on his or her performance or outcomes in a manner that accords with what is found offline (Sections 5.2.2 and 9.3.2).

1.6 SHAPING FORCE OR SOCIAL TOOL?

The final consideration that helps to provide context for this book is the question of whether the web has changed behaviour or is more a tool that people use to pursue their social, economic and political ends. While there

is no doubt that the web has had (and is continuing to have) a remarkable impact on the world, and many researchers focus on understanding how the web transforms behaviour, the focus of this book is more on the latter question.

In considering the impact of the web (and, in particular, the concept of a virtual community), Fischer (1997) drew on his previous research findings that the influence of new technologies on patterns of communication and community was moderate, in comparison to other factors such as demography and economic forces. This led Fischer to conclude that 'we ought to think more about [new technologies] as tools people use to pursue their social ends than as forces that control people's actions' (p. 115).[7]

The present book is not focused so much on the web as a transformative technology but rather as a technology that people make use of in their social, economic and political behaviour. The book is focused on types of behaviour that have been studied by social scientists for a long time, but identifies the opportunities and challenges that are presented by digital social data.

1.7 CONCLUSION

This chapter has provided an introduction to the key technologies that underlie the web and has outlined some of the major events and phases in the development of the web. Prominent examples of online computer-mediated interaction that feature elsewhere in the book were introduced. The chapter also aimed to provide an introduction to web social science, showing how it differs from other academic approaches for studying the web. This was done by first outlining three key phases in how people have conceptualised the web: cyberspace, virtual communities and online social networks. It is the latter approach (online social networks) that is most relevant to this book.

Another way of distinguishing web social science is to look at various disciplinary approaches to studying the web, and this chapter identified four such approaches: network science (as studied by applied physicists and computer scientists), information science, media studies and network science (as studied by social scientists). It was argued that the social science approach to network science is distinct from the other three approaches. Web social science (as presented in the remainder of the book) draws mainly from social scientists' perspective on network science, although contributions from applied physics and information science also feature.

Finally, it was noted that a key distinguishing feature of this book is the perspective that, rather than being a force that is shaping human behaviour,

[7]It would be interesting to see whether Fischer's conclusion would be different today, since 1997 was early in the history of the web.

the web can perhaps best be viewed as a tool that people use to achieve social, economic and political outcomes. The web provides social scientists with a unique data source for studying this behaviour, thus providing new insights into long-standing questions in social science.

Further reading

Flew (2008) provides an introduction to Internet law, policy and governance. See Bruns (2008) for more on prosumption and produsage. See Rheingold (2000) for more on virtual communities.