Chapter 4

Hyperlink networks

Last revision: 29May2012

This chapter provides an introduction to hyperlink network research, with the main focus being on Web 1.0 websites (or the "static" Web) (Box 1.3).[1]

When we think of Web 1.0, we are typically focusing on organisations (e.g. corporations, government agencies, NGOs) who are maintaining websites, and in Section 4.1 there is a discussion about why organisations create hyperlinks and the benefits of receiving hyperlinks. Section 4.1 also introduces three fundamental questions about defining a hyperlink network (what is the tie?, what are the nodes?, where are the boundaries?), drawing on the related discussion in the context of social networks from Section 3.1. Section 4.2 outlines three disciplinary approaches to social scientific research into hyperlink networks: hyperlink networks as citation networks (library and information sciences), hyperlink networks as issue networks (media studies) and hyperlink networks as social networks (sociology). In Section 4.3 there is an introduction to tools for hyperlink retrieval (web crawlers) that have been used in humanities and social science research.

4.1 Hyperlink networks – background

The Web is a vast electronic library of hyperlinked documents, but the social scientific approach to studying hyperlinks involves conceptualising the Web as being something other than simply a repository of electronic information.[2] As Jackson (1997) noted, "Once we become critical of the assumption that the Web is a neutral repository of information, the structure of the Web becomes much more interesting."

However, in order to study the structure of the Web from a social scientific perspective, it is important to first think about why organisations create hyperlinks (and the benefits they gain from receiving hyperlinks) and also we need to consider the key methodological question of how to construct a hyperlink network for research purposes.

4.1.1 Motives for sending and benefits of receiving hyperlinks

Social science web research often involves studies of groups, organisations or companies where we are trying to learn something about what it means to receive or send a hyperlink from the perspective of

[1] The technology that underlies the Web 1.0 website is very similar to that of webblogs, and so there is also a discussion about the latter in Section 4.3.3

[2] In this chapter references to the Web are to be taken as meaning "Web 1.0".

organisational behaviour. In the context of unobtrusive (nonreactive) research, where we are using a web crawler to collect hyperlink data without asking the website owners why they created the link or looking at the text surrounding the hyperlink, it is not straightforward to know what is being exchanged by a hyperlink.

Sending hyperlinks

The hyperlink is commonly seen as the "essence" of the Web (Jackson, 1997; Foot et al., 2003). There are several possible interpretations that have been offered for why a hyperlink might be created i.e. why hyperlinks are *created* or *sent*. Hyperlinks can be seen as "conferrers of authority" or endorsement (Kleinberg, 1999), indicators of trust (Davenport and Cronin, 2000), reflections of organisational communicative and strategic choices (Rogers and Marres, 2000), and as tools of organisational alliance building and message amplification (Park et al., 2004).

The many interpretations of the meaning of a hyperlink between two websites has led Thelwall (2006) to conclude that there can be no single "Theory of Linking". Certainly, the motivation for hyperlinking will vary depending on the context – a government department, for example, will have a different motivation for hyperlinking than a political party.

In Section 6.2.1 we discuss Shumate and Dewitt (2008) who conceptualise the hyperlinking activities of NGOs focused on HIV-AIDS as being directed towards the creation of an information public good (a hyperlink network that enables information on this issue to be located). In contrast, other researchers (Ackland and O'Neil, 2011) regard the hyperlinking behaviour of activist organisations as being related to the formation of online collective identity (Section 6.3). In Section 8.1 the number of outbound links from a government department website is seen as a measure of "extroversion", providing a quantitative measure of the prominence or centrality of the website in social and information networks, also known as "nodality" (Hood, 1983; Hood and Margetts, 2007).

Receiving hyperlinks

With regards to the *receipt* of hyperlinks, as discussed further in Section 7.1.1, inbound links are important in driving traffic to websites for two reasons. First, the more inbound hyperlinks from other relevant websites, the greater the number of "pathways" that people can follow to the website. Second, inbound hyperlinks are a primary determinant of a site's ranking on search engines such as Google. As put by Hindman et al. (2003) website *retrievability* is an absolute concept (if the website is "down" the content is not viewable, but as long as it is "up" it is as viewable as the content on any other website). Website *visibility*, however, is relative, and is largely determined by the number of inbound hyperlinks from other relevant websites.

In Section 7.1, the importance of receiving hyperlinks is discussed in the context of the visibility of political information and the question of whether counts of inbound hyperlinks can be used as scientometric measures of academic authority and output is addressed in Section 9.2.1.

4.1.2 Hyperlink network nodes, edges and boundaries

As noted in Section 3.1, the fundamental methodological challenge in social networks research is defining the nodes, ties and boundaries to the network. We now look at each of these in the context of hyperlink networks.

Hyperlink network nodes

Defining the nodes in a hyperlink network can be more complicated than with offline social networks, and even other types of online networks. In a real-world friendship network, the nodes are obviously people, and the nodes in Facebook and Twitter networks are similarly easy to define, since each username in these social media environments is typically associated with a particular individual.

However Web 1.0 hyperlink networks may be populated by nodes that are not homogenous in type. For example, it is very easy to construct a hyperlink network where nodes will represent organizations such as universities, government departments, companies, non-government organizations (NGOs) - these are the websites of entities that have an offline or real-world presence. But additionally, there may be nodes that represent entities that have no offline presence i.e. they only exist on the Web. Examples are: portals (sites that provide organized lists of links to other sites), informational sites (sites that aim to provide commentary or insight on a particular topic and links to relevant resources), online businesses (that have no offline counterpart), websites operated by individual people, vanity websites (websites that have been set up to promote a brand, or movie) and the popular special form of web pages known as blogs.

In some research contexts, it may make sense to prune the hyperlink network so that it only contains nodes representing organizations that exist offline. In other situations it may be desirable for the hyperlink network to be an accurate representation of what exists on the web i.e. to include all websites found by the web crawler, regardless of their type. Even if a decision is made to just focus on websites that represent organisations with an offline presence, there are additional methodological challenges that are particular to hyperlink network analysis.

Ideally, the nodes in an organizational hyperlink network will typically represent the entire (Web 1.0) web presence of each organization, that is, websites, or parts of websites, rather than individual web pages. Hence the hyperlink network will need to be populated with meta-nodes representing all web pages in a site, rather than dozens or perhaps hundreds or thousands of nodes reflecting the individual pages in each site. This can be achieved algorithmically by grouping together all pages from a particular hostname. However, sometimes an organization's web presence may be reflected in multiple hostnames e.g. different domain names reflecting, for example, international business presence (www.mycompany.com, www.mycompany.com.au) or different sub-domains (brand1.mycompany.com, brand2.mycompany.com) or different sub-sites (e.g. www.mycompany.com/brand1, www.mycompany.com/brand2).[3] In order to accurately measure the web presence of the entire organization all the identified pages from these domains, sub-domains and sub-sites will need to be properly grouped together. While a lot of this processing can be done automatically, there will typically still need to be some input from the analyst in order to accurately identify the nodes in the hyperlink network that should be grouped together.

Hyperlink network ties

The second methodological question to answer before conducting hyperlink network analysis is: what are the network ties? Assume the simplest case of two organizations whose web presence is reflected in single websites. Even in this simple situation there are several possibilities as to what could constitute a network tie. If the website of organization A contains a hyperlink to the website of organization B, then this could be interpreted as a network tie and it would lead to creating a directed edge in the hyperlink network. But you might be more interested in only recording a network tie if the hyperlinks are reciprocated (A links to B, and B links to A), leading to an undirected, symmetric hyperlink network.

[3] Note that sub-sites can cause a further problem: it may be that two or more organizations in your hyperlink network have websites that are commercially hosted (for example by Geocities) and in such a situation, you need to be careful that these websites are not merged into a single node e.g. www.geocities.com.

Finally, you might want to attach values or weights to the network ties reflecting, for example, the number of hyperlinks directed from A to B (one could argue that more hyperlinks reflect a stronger network relationship) or the depth in the website where the hyperlink was embedded (one could argue that a hyperlink from the homepage of A has more significance than if it is buried deep within the website).[4]

Since sending and receiving hyperlinks can mean different things depending on the organisation (Section 4.1.1) it is important that hyperlink networks be constructed such that each hyperlink can be interpreted in the same way. The assumption here is that research is being conducted in an unobtrusive manner (digital trace data) and at a scale that it is not feasible for the researcher to go to each page where there is a hyperlink and check the page content to ascertain the context or meaning of the hyperlink. In the absence of manual checks into the context or meaning of a hyperlink, a straightforward way of ensuring this consistency in the meaning of hyperlinks is to only include particular types of network actors. Lusher and Ackland (2011) construct a network of websites of organizations who advocate on behalf of refugee and asylum seekers in Australia. The fact that the network was restricted to these types of actors meant that the authors could interpret the existence of a hyperlink using a collective action framework i.e. the actors were hyperlinking to other organizations who shared their concern for the plight of refugees and asylum seekers in Australia and are working towards this cause. Lusher and Ackland purposefully excluded government sites from the network since the existence of a hyperlink from a refugee advocacy site to a government site could reflect criticism of government policy, and the selection of nodes thus ensured that hyperlinks were likely to represent positive affect.

Hyperlink network boundaries

The final methodological question to address is: what are the boundaries to the hyperlink network? Network boundaries will often be settled by choice of type of node, but a complication with constructing hyperlink networks (compared with offline networks) is that it might not be possible to use familiar boundaries like geography to define the limits of a web hyperlink network. In an offline friendship network the boundary might be the school (or classroom) i.e. if a student has a friendship with someone outside the school, then this friendship will not be counted. However such geographical or physical boundaries do not necessarily exist or make sense in the "borderless" terrain of the Web.

However, in some situations, geography might be helpful in determining hyperlink network boundaries. Lusher and Ackland (2011), for example, only included the websites of Australian-based organizations who are involved in refugee and asylum seeker advocacy. However even in such a situation where an obvious network boundary can be used, the analyst may still be faced with the problem of discovering all of the units of observation (nodes) that exist within the boundaries and (if necessary) drawing an appropriate sample of nodes. Lusher and Ackland used a snowball sampling approach (Section 2.2.1) to discover relevant websites - starting with an initial list of known relevant sites and then crawling these in an attempt to find additional relevant sites - but unless the analyst has available a sampling frame this process is problematic as you will never know if you have included all relevant websites into the hyperlink network.

[4] Alternatively, it may be that a link directed to a page that is deep within the website is valued more highly than a generic link to a top-level page.

4.2 Three disciplinary perspectives on hyperlink networks

In this section, we introduce three disciplinary perspectives for the quantitative study of hyperlink networks.[5]

4.2.1 Citation hyperlink networks

Webometrics is a collection of techniques for quantitatively measuring documents and information from the Web, and has its disciplinary origins in informetrics, which is a sub-field of library and information science (Section 1.4). Thelwall (2009b) notes that webometrics has four main areas (text content analysis, analysis of hyperlink structure, web usage analysis and web technology analysis), but here, we focus on webometric contributions to hyperlink analysis.

In Section 9.2.1 we discuss an example of webometric hyperlink research, Barjak and Thelwall (2008), who study the factors associated with the prominence or visibility of websites belonging to research teams in the biological sciences. In this study, counts of inbound links are regressed on the characteristics of the site and site owner, in order to identify those qualities that influence inbound links. The study aims to assess the viability of hyperlinks as scientometric performance indicators, and it highlights the fact that webometric hyperlink research involves the conceptualisation of inbound hyperlinks as being akin to citations that are traditionally studied by informetricians.

However, it should be noted that webometric techniques have been used in areas outside of scientometric research. For example Margolis et al. (1999) and Gibson et al. (2003) focus on counts of inbound hyperlinks to minor and major political party websites in their test of the "normalisation hypothesis" (Box 7.1). The use of counts of inbound hyperlinks to and outbound hyperlinks from government department websites in establishing the "nodality" of different government department websites by Escher et al. (2006) is also an application of webometrics (Section 8.1).

4.2.2 Issue hyperlink networks

The concept of issue networks (see, e.g. Rogers, 2010a,b) is an example of studying hyperlinks from the media studies (Section 1.4) perspective. Issue networks emerge when actors who are engaged in a common issue generate an "associational space", which is defined by hyperlinks. According to Rogers (2010), issue networks are the representation of public debate (the initial concept was a circle diagram depicting actors sit at a virtual "table" to discuss a particular issue), where the network nodes can be people and organisations, but also "argument objects" such as news items, documents, or any type of content that is relevant to the issue. The conceptualisation of issue networks thus draws on Actor Network Theory (Latour, 2005), envisaging actors that are both material and conceptual.

issue network example

4.2.3 Social hyperlink networks

The potential for using social network analysis (Chapter 3) to analyze hyperlink networks was noted by Jackson (1997) who considered that SNA "...has significant potential to generate insight into the communicative nature of Web structures". But Jackson (1997) was not comfortable with the idea of nodes in a hyperlink network (pages or sites) being described as social actors and the core assumption of SNA – the inter-dependence of nodes within a network – as being applicable to the Web.

[5] This section draws on Borquez and Ackland (2012).

While Jackson (1997) not sanguine that formal SNA concepts and methods could carry over to the Web, other authors have had less reservations and Park (2003) advocated that the analysis of hyperlink networks using SNA be called "hyperlink network analysis". However, despite this early recognition of the potential of SNA for hyperlink analysis, there are not many examples of formal SNA techniques being used to analyse hyperlink networks.

Notable exceptions are Shumate and Dewitt (2008), Gonzalez-Bailon (2009), and Ackland and O'Neil (2011) (see Chapter 6) who use ERGM in the context of analysing organisational collective behaviour on the Web. Lusher and Ackland (2011) refer to the application of ERGM to hyperlink networks as *relational hyperlink analysis* and show that this approach can provide fundamentally different conclusions about the social processes underpinning hyperlinking behavior, compared to hyperlink counts regressions (which is a hallmark of webmetrics). In particular, counts regressions may over-estimate the role of actor attributes in the formation of hyperlinks when endogenous, purely structural network effects are not taken into account.

4.2.4 Comparing the disciplinary perspectives

Having introduced the three disciplinary approaches to conceptualising hyperlink networks (citation hyperlink networks, issue hyperlink networks and social hyperlink networks), we now attempt to show how they differ. The network definition triumvirate (Section 3.1)–(who are the nodes?, what are the edges? and where are the boundaries?)–can be used to elucidate differences between the three approaches to hyperlink network research. Other useful dimensions on which to compare the three approaches to conceptualising hyperlink networks are network type (Section 3.1) and the unit of analysis.

Who or what are the nodes?

For citation hyperlink networks and social hyperlink networks, the nodes are typically websites that represent groups/organizations (that may or may not have an offline presence) or individuals (e.g. blogs). With issue hyperlink networks, as noted above, the nodes can either represent social actors or objects (e.g. news items, documents) that are relevant to the issue being studied.

What constitutes a tie?

At one level, this question seems quite simple because a hyperlink has a clear definition in terms of the technologies (HTML and HTTP) that allow one to connect documents across the web. But what meaning do we ascribe to the existence of hyperlink?

As with the question above about who are the nodes, there are similarities in what constitutes a tie in citation hyperlink networks and social hyperlink networks. Typically, hyperlinks in such networks indicate positive, rather than negative affect relations. That is, the hyperlink is likely to be transferring symbolic or practical resources that the recipient will want to receive. For example, in Barjak and Thelwall (2008) hyperlinks to academic project websites are conveying practical resources in the form of intellectual authority, status or prominence Section 9.2.1. In Ackland and O'Neil (2011), the hyperlinks between environmental activist websites are modelled as conveying both practical resources ("index authority") and symbolic resources ("boundaries of belonging"), which help to establish online collective identity Section 6.3.

As noted by Lusher and Ackland (2011), there is a point of departure between citation hyperlink networks and social hyperlink networks in terms of the nature of the tie. Webmetrics is generally focused on the characteristics of actors that lead to the *acquisition* of hyperlinks ("actor-relation

effects", using ERGM terminology – see Section 3.2.2). Even though they used regression rather than ERGM, Barjak and Thelwall (2008) are effectively identifying a particular type of actor-relation effect, receiver effects, which are the website attributes that are associated with a tendency to receive hyperlinks. In contrast, social hyperlink network research involves identifying all possible social forces that have led to the emergence of a given hyperlink network. ERGM is particularly well suited to this task, since it can be used to identify both actor-relation effects and hyperlinks that are unrelated to actor attributes, representing more informal networking that occurs between social actors because of social norms such as reciprocity (i.e. endogenous or structural network effects).

Once again, when it comes to the nature of the tie, the issue network is quite different to the social hyperlink network and citation hyperlink network. An issue network contains any actor with a stake or involvement in an issue and for this reason, we can expect the existence of negative affect relations in issue networks. An environmental issue network might contain environmental activists, government and business actors and it is likely that hyperlinks from activists to business (and possibly government) will represent negative affect relations. In contrast, the environmental activist networks studied by Ackland and O'Neil (2011) only contained activists. In their study of refugee and asylum seeker advocacy networks, Lusher and Ackland (2011) purposely excluded government websites because their presence would make it harder to interpret the meaning of hyperlinks due to the fact that there would be a mix of negative and positive affect relations.

Finally, because issue networks can contain nodes that represent objects (news items, documents) that pertain to an issue, it is difficult to conceive of an issue network as containing either actor-relation or purely structural effects. So again, this is where issue networks differ from the other two types of hyperlink networks.

Where is the network boundary?

In the case of citation hyperlink networks, one would think that the network boundaries are generally going to be clear in the sense that the researcher will have a list of entities that exist in the real world (e.g. research teams, university departments, universities) and these entities are what comprise the actors of the network. Either all the entities will be included in the analysis (i.e. a census) or else a sample of the entities. But it needs to be realised that citation hyperlink network analysis generally involves the construction of counts of inbound links from *anywhere on the Web*, not just from actors included in the analysis. So in that sense, in the case of citation hyperlink network analysis the network boundaries are not clear cut as one might at first think.

In the case of social hyperlink networks, in some situations, the network boundaries might be clear. For example, a study of the hyperlink networks of environmental activists might focus only on those organisations that have an offline presence i.e. activist organisations that are registered (for tax purposes) as non profits working on the environment. In such a situation the researchers could decide whether to conduct a census (include all such organisations in the study) or draw a random sample of organisations. However, with many examples of social hyperlink network research the underlying population of relevant websites cannot be identified in advance since the relevant population might be, for example, all websites run by organisations that are focused on a particular issue. In such a case, a snowball sampling approach is needed to draw additional nodes into the network.

With an issue network, as with many examples of social hyperlink networks, the underlying population of relevant websites cannot be identified in advance and a snowball sampling approach is needed to draw additional nodes into the network.

Network types

We can use network types (Section 3.1) to further distinguish citation, social and issue hyperlink networks. A citation hyperlink network is in fact a series of 1 degree egonetworks, with the focal nodes being the sites of interest. In contrast, social and issue hyperlink networks will generally be complete networks.

What is the unit of analysis?

The main technique used to analyse citation hyperlink networks is regression, and hence the unit of analysis is the website. In contrast, social hyperlink networks are analysed using ERGM and thus the unit of analysis is the dyad (pairs of websites). With issue networks, the unit of analysis is any object or actor that is connected to the issue.

4.3 Tools for hyperlink network research

This section provides an introduction to webcrawlers and also discusses sources for historical web data. A particular type of web page, blogs, is also introduced.

4.3.1 Webcrawlers

A web page contains two types of data that are of interest to us, hyperlinks and text content, which are embedded in the HTML content in the page. While it is possible to use a web browser to collect these data from web pages, such an approach is very time consuming and clearly not feasible for large numbers of websites (or indeed for websites that contain a lot of pages). For some time now researchers have been using **web crawlers**, which are software tools that automatically traverse a web site by first retrieving a single web page (for example, entry or top-level page on a site) and then recursively retrieving all web pages that are referenced (e.g. following hyperlinks throughout the site). A web crawler can save a copy of each web page that it encounters, but it can also parse the HTML content, extracting the hyperlinks and text content. In order to automatically extract data from web pages, web crawlers rely on the web pages being written to comply with the HTML specification (Box 4.1).

Three webcrawlers used in the social sciences

The following are three examples of publicly-available web crawlers that are used for social science research.

IssueCrawler[6] is web-based software for hyperlink network construction and analysis, developed under the direction of Richard Rogers. The first version of Issuecrawler was released in 2001 and hence it is a pioneering example of a research tool that is accessible via a web browser (such tools are increasingly common, with the movement of services such as email and productivity tools into the "cloud"). With its origins in media studies (although, widely used in the social sciences), Issuecrawler has been primarily designed as a web crawler for the construction and visualisation of issue hyperlink networks (Section 4.2.2).

Starting with a list of seed URLs (e.g. web pages focused on a particular issue), Issuecrawler crawls the seed set using three different approaches (Govcom.org, 1995). First, "co-link analysis" is where only

[6] Govcom.org, `http://www.issuecrawler.net/`

The following is the HTML for the homepage of a fictitious travel website (http://www.robstravel.com):

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<HTML>
   <HEAD>
      <TITLE>My first HTML document</TITLE>
      <META name="keywords" content="travel service,holidays,Australia">
      <META name="title" content="Rob's Travel">
      <META name="description" content="We specialise in holidays to Australia!">
   </HEAD>
   <BODY>
      <P>Welcome to Rob's Travel!</P>
      <P>You may also like to see the website of our partner
      <A href="http://www.robsrentalcars.com">Rob's rental cars</A>, who rent cars.</P>
   </BODY>
</HTML>
```

The meta keywords and description provide information about the nature of the site, while the page body contains a hyperlink to a partner website (http://www.robsrentalcars.com). A webcrawler can extract the meta data, body text and hyperlinks, but social science hyperlink research will often require manual coding of the site and possibly even hyperlinks. It has been suggested that the Semantic Web might obviate the need for manual coding of sites and hyperlinks, since the framework involves the use of markup languages that allow website owners to present basic information about the site in a machine-readable manner. To illustrate, the above example could be "semantified" with the addition of the following RDF document:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="http://www.robstravel.com/rdf/">
<rdf:Description rdf:about="http://www.robstravel.com">
  <si:title>Rob's Travel</si:title>
  <si:product>travel service</si:product>
  <si:category>holiday</si:category>
  <si:country_specialisation>Australia</si:country_specialisation>
  <si:partner>Rob's Rental Cars</si:partner>
  <si:partner_url>http://www.robsrentalcars.com</si:partner_url>
  <si:partner_product>rental cars</si:partner_product>
</rdf:Description>
</rdf:RDF>
```

Rob's Travel semantically-enabled website would use an *ontology* (an exact description about objects and their relationships) providing precise meanings of the tags "product", "category", etc. As long as the webcrawler understands the ontology, then the site could be automatically coded as a business that is providing travel services, specialising in holidays to Australia, which is affiliated with a rental car company.

However, Brent (2009) contends that the Semantic Web may not be a boon for social science Web researchers for two main reasons. First, the human effort involved with making a website both human- and machinable-readable (i.e. presenting the information in both HTML and RDF) is significant, and while there might be incentives for this to happen in particular domain areas (e.g. libraries and e-commerce) it is less likely to occur in areas of interest to social scientists e.g. advocacy and protest sites. Second, it is questionable that a single ontology can adequately capture the diversity of parts of the Web that are of interest to social scientists. While one can envisage real-world examples of ontologies successfully being used to make e-commerce sites machine-readable, it is hard to imagine organisations participating in, for example, the abortion debate agreeing to and then implementing an ontology that would allow social scientists to automatically code their websites on the basis of their stance on abortion.

Box 4.1: HTML and RDF

those web sites receiving at least two links from the seed set are included into the final network. Of note is the fact that a seed site will be excluded from the final network if it does not receive links from at least two other seed sites. Co-link analysis in Issuecrawler needs to be differentiated from how the term is used in information science, where it refers to the construction of a tie between two entities (e.g. articles, authors) because they both link to a third entity (even if they do not directly link to one another). In Issuecrawler, co-link analysis simply refers to the method by which websites are selected to appear in the final network, with links between websites in the final network being hyperlinks.

The second technique for constructing a network in Issuecrawler is "snowball analysis" which is equivalent to snowball sampling (Section 2.2.1): all of the outbound links from the seed sites are included into the network and these sites are then themselves crawled (this process continues up to three degrees of separation from the seed set). Finally, "inter-actor analysis" displays interlinking between the seed set exclusively, that is, it allows the construction of a complete network.

SocSciBot[7] is a long-established web crawler and hyperlink network analysis tool that was developed by Mike Thelwall (Thelwall, 2009b). Unlike Issuecrawler, SocSciBot is client software (you download and install on your own computer).

Mike Thelwall is one of the pioneers of webometrics and hence SocSciBot was developed for the construction and analysis of citation hyperlink networks (Section 4.2.1); the earlier versions of SocSciBot were primarily designed for measuring web impact through retrieving counts of inbound links and did not include a network visualisation tool. However, as with Issuecrawler, SocSciBot can be used for other types of hyperlink research. For example, Shumate and Dewitt (2008), which we regard as an example of social hyperlink network research, employed SocSciBot for the data collection, and the current version of SocSciBot provides network visualisation and is designed to be used in combination with the social network analysis tool Pajek.[8]

Virtual Observatory for the Study of Online Networks (VOSON)[9] is a tool for collecting and analysing hyperlink network data and like IssueCrawler, it is a hosted service available via a web browser. VOSON was first publicly released in 2006 and it has been designed specifically for the analysis of social hyperlink networks (e.g. Ackland and Gibson, 2004; Ackland, 2009, 2010b; Lusher and Ackland, 2011).

APIs

SocSciBot, IssueCrawler and VOSON all feature web crawlers as their main data collection tool. However, often we want to know not just the hyperlinks that are being directed *from* a given website (i.e. outbound hyperlinks), but also what hyperlinks are being directed *to* the website (i.e. inbound hyperlinks). Search engines such as Google and Bing both allow you to find this information manually via their search engine websites: if you put "link: voson.anu.edu.au" into Google, it will list all of the web pages that hyperlink to the VOSON website. However, for large-scale research, the Google and Bing search sites are not useful, and instead it is possible to use application programming interfaces (APIs) to enable software to query the databases directly.[10] For more on APIs in the context of hyperlink research, see Thelwall (2004, 2009b).

[7] Statistical Cybermetrics Research Group, University of Wolverhampton, `http://socscibot.wlv.ac.uk/`

[8] `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

[9] VOSON Project, The Australian National University, `http://voson.anu.edu.au`. It should be noted that the author created the VOSON software and is involved in the commercial development of the software.

[10] See `http://code.google.com/` and `http://www.bing.com/toolbox/bingdeveloper/`

Ethics of using webcrawlers

In addition to points made in Section 2.8, there are particular ethical issues associated with the use of webcrawlers for research (Thelwall and Stuart, 2006). First, crawling a website can potentially use a lot of the resources (e.g. bandwidth, CPU time) of the website owner, which could lead to significant costs or loss of service quality. For this reason it is important that webcrawlers are used responsibly, for example by not crawling the sites of organisations who might be resource constrained (e.g. NGOs in developing countries), and also by limiting the crawler so there are delays between each page request. Second, it is important that webcrawlers obey the robots.txt protocol[11], which is used by webmasters to inform crawlers which parts of the website can be crawled and which parts are "off limits".

4.3.2 Historical web data

While the above tools can be used to collect "live" website data, what about if we want to study how a group of websites have changed over time?[12] For example, say we were interested in using Web data to look at how climate change has emerged as an issue of concern, and how various actors (government, corporate, NGO) are responding to climate change. To conduct such research we either need access to historical web data or else we need to have been periodically crawling the web ourselves, thus constructing a time series of webcrawls (and network datasets).

The Internet Archive[13] was founded in 1996 as an Internet library, offering "permanent access for researchers, historians, and scholars to historical collections that exist in digital format", and it is a member of the International Internet Preservation Consortium (IIPC) (Box 4.2) which also includes many national libraries as members. The Internet Archive has been crawling the web since 1996 and currently, the archived web pages are publicly available via the Wayback Machine, a browser interface where if you enter a URL of a website, you will see the points in time when the the website was crawled and archived. There is currently no way to automatically extract hyperlinks or text content from websites that have been archived by the Internet Archive. This means that while the Wayback Machine is fine for qualitative or descriptive research involving a single website or a small number of websites, it is not suitable for large-scale empirical research. It would be impossible, for example, to study the evolution of the hyperlink networks formed by a couple of hundred environmental activist websites. However, the Internet Archive is currently developing an API into its archive that will eventually (hopefully) facilitate automated querying of their database in a manner that is currently possible with Google and Bing.

4.3.3 Blog sites

The underlying technology of a blog site is the same as a website: a blogsite is simply a chronologically updated website, generally authored by a single person, with a diary or commentary style. Web crawlers (such as those used by SocSciBot and VOSON) can access blog sites to extract hyperlinks and text. However, the structure of blog pages makes the data collection via web crawlers more challenging than is the case with static websites, for two reasons. First, a standard web crawler will not be able to distinguish between **permalinks** (these are hyperlinks that appear within a given blog post), **blogroll** links (these are the hyperlinks that often sit to the side of the page i.e. they aren't contained within a particular blog post), and links that might appear in the comments section on the blog page (i.e. links not made by the blogger, but by someone else commenting on the blog post). Permalinks (contained in

[11] `www.robotstxt.org/`

[12] Note that data collected via the Bing and Google APIs will not necessarily be current, since the data is extracted from the databases of these organisations, however they are continually crawling the web so we can expect the data will be reasonably up-to-date.

[13] `http://www.archive.org`

The International Internet Preservation Consortium (IIPC)[14] was formed due to the recognition of "... the importance of international collaboration for preserving Internet content for future generations." The goals of the consortium are (as listed on their website):

- To enable the collection, preservation and long-term access of a rich body of Internet content from around the world.
- To foster the development and use of common tools, techniques and standards for the creation of international archives.
- To be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content.
- To encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation

One of the main objectives of the IIPC has been to develop open-source tools for setting up a web archiving "chain" (this is the workflow involving the collection, storage and access of web materials). The IIPC working groups have developed the following tools[15]:

- **Acquisition. Heritrix** is an open-source, extensible, Web-scale, archiving quality Web crawler. This is the crawler used by the Internet Archive.
- **Curator Tools.** These are tools to enable librarians to define and control harvests of web material. Examples are: **Web Curator Tool (WCT)** and **NetarchiveSuite**.
- **Collection storage and maintenance.** Heritrix stores web crawls using a file specification called "ARC". **BAT (BnFArcTools)** is an API for processing ARC files.
- **Access and finding aids. Wayback** is "a tool that allows users to see archived versions of web pages across time" (this tool is the basis for the Internet Archive's Wayback machine). **NutchWAX** (Nutch with Web Archive eXtensions) is "a tool for indexing and searching Web archives using the Nutch search engine and extensions for searching Web archives". **WERA (WEb aRchive Access)** is "a Web archive search and navigation application. WERA was built from the NWA Toolset, gives an Internet Archive Wayback Machine-like access to Web archives and allows full-text search."

Box 4.2: International Internet Preservation Consortium

blog posts) are generally made by the blogger as he/she comments or points to other blog posts, or web pages on traditional media sites. Permalinks thus generally reflect the current "reading" behaviour of the blogger, and are often considered by researchers to be a more accurate indicator of the links that bloggers are making to one another and to other sites. In contrast, blogroll links generally reflect more permanent affiliations between bloggers; for example, it is common for political bloggers to have blogroll links to other bloggers who share the same political persuasion. Blogroll links can become "stale" and thus of less value/importance in blog analysis.

The second challenge faced in collecting data from blog sites relates to the chronological structure of the average blog page. Generally blog sites are structured so that all the blogs posts for a given month appear on the same page (with most recent posts at the top of the page). One of the main aims of blog research is to identify links that are made within a given time period - this allows the tracking of the diffusion of influence throughout the blogosphere, for example (where was an issue first taken up, and how did it spread throughout the blogosphere?). In order to get this type of dynamic data, it is necessary to ensure that the blog pages are parsed up so that links can be attributed to a particular time period.

The above challenges regarding the collection of blog data are not insurmountable, but researchers are advised to make use of specialised services for providing data from the blogosphere. An example of such as service is the Blog Analysis Toolkit[16], which is a hosted webcrawler specifically designed for blogs, and Infoscape[17] have also developed tools for extracting data from the blogosphere. There are also several APIs for blog data; Ackland (2005) used Bloglines[18] which was an early blog API, and Spinn3r[19] provide a blog API and have also released several large blog datasets to the research community (Burton et al., 2009).

4.4 Conclusion and further reading

This chapter began with a discussion of various motives for creating hyperlinks and also the benefits an organisation may gain from receiving a hyperlink. We then discussed the definition of a hyperlink network, focusing on how we can understand nodes, edges and boundaries to networks in the context of hyperlink networks. The chapter then proposed three broad disciplinary approaches for empirically studying hyperlink from a social science perspective, where hyperlink networks can be conceived of as citation, issue or social networks. The final section was devoted to tools for collecting (and analysing) hyperlink data, with focus on three webcrawlers that are used for hyperlink research: Issuecrawler, SocSciBot and VOSON.

Further reading

For more on webometrics and the use of webcrawlers in the social sciences more generally, see Thelwall (2004, 2009b) and also the SocSciBot website.[20] The Digital Methods Initiative website[21] and Govcom.org website[22] provide more on the concepts and methods relating to issue networks and the Issuecrawler software. Finally, more information about the VOSON software and related methods can be found on the VOSON project website.[23]

[16] https://surveyweb2.ucsur.pitt.edu/qblog/page_login.php
[17] http://www.infoscapelab.ca
[18] http://www.bloglines.com
[19] http://www.spinn3r.com
[20] socscibot.wlv.ac.uk
[21] http://wiki.digitalmethods.net/Dmi/WebHome
[22] http://govcom.org
[23] http://voson.anu.edu.au