1

Analysing the UK web domain and exploring 15 years of UK universities on the web

Eric T. Meyer, Taha Yasseri, Scott A. Hale, Josh Cowls,
Ralph Schroeder and Helen Margetts

Introduction

The World Wide Web is enormous and in constant flux, with more web content lost to time than is currently accessible via the live web. The growing body of archived web material available to researchers is potentially immensely valuable as a record of important aspects of modern society, but there have previously been few tools available to facilitate research using archived web materials (Dougherty and Meyer, 2014). Furthermore, based on the many talks we have given over the years to a variety of audiences, some researchers are not even aware of the existence of web archives or their possible uses. However, with the development of new tools and techniques such as those used in this chapter and others in this volume, the use of web archives to understand the history of the web itself and shed light on broader changes in society is emerging as a promising research area (Dougherty et al., 2010). The web is likely to provide insight into social changes just as other historical artefacts, such as newspapers and books, have done for scholars interested in the pre-digital world. As the web becomes increasingly embedded in all spheres of everyday life and the number of web pages continues to grow, there is a compelling case to be made for examining changes in both the structure and content of the web. However, while interfaces such as the Wayback Machine[1] allow access to individual web pages one at a time, there have been relatively few attempts to work with large collections of web archive data using computational approaches across the corpus.

The research presented in this chapter used hyperlink data extracted from the Jisc UK Web Domain Dataset (Jisc, n.d.-a) covering the period from 1996 to 2010 to undertake a longitudinal analysis of the United Kingdom (UK) national web domain, .uk, focusing on the four largest second level domains: .co.uk, .org.uk, .gov.uk, and .ac.uk. We explore the growth of these domains, and examine the link density within and between them. Next we look in more detail at the academic second-level domain, .ac.uk, to understand the relationship between link density among UK academic institutions and measures of affiliation, status, performance and geographic distance. Overall, these results are used both to understand the growth and structure of the .uk domain, but also to demonstrate the benefits and challenges of this type of analysis more generally.

Background

Archiving national web domains

National web domains represent one approach to web archive analysis for researchers seeking an overview of a single country's web presence (Brügger, 2011). Any particular national web domain offers the potential of both diversity and completeness in its coverage (Baeza-Yates et al., 2007), although there are limitations in terms of generalizability beyond the country in question and frequently in terms of the completeness of the analysis based on technical factors (see section on the UK web domain below). At the same time, limiting the focus to a single country reduces the number of contextual differences (such as multiple dominant languages, different internet and broadband penetration rates, different degrees of political openness and so forth), and thus is a sound strategy for demonstrating the potential of this new type of analysis.

Research in this area is at an early stage, and there are conceptual challenges associated with analysing national web domains. The content and structure of country-code top-level domains (ccTLDs), such as .uk for the UK and .fr for France, are governed more by tradition than rules (Masanès, 2006), complicating efforts to reach a comprehensive definition of what they represent. Brügger (2014) discusses the difficulty, for example, of deciding how national presences should be delimited. In the case presented here, the domain name .uk is used, but this does not cover all the web pages originating in the UK as it is possible for UK companies, organizations and individuals to

use generic top-level domains (.com, .org, etc.) or those assigned elsewhere. Moreover web pages ending with .uk are also used for websites which arguably belong to a different country, as when multinational companies headquartered outside the UK have affiliates within the UK with a .uk address. Finally, it might be contended that not only web pages with a .uk address be examined, but also those that link to and from these web pages. However, for the purposes of this research, these limitations can mostly be noted for future research and do not seriously limit the ability to understand the broad patterns within the UK national web presence. Furthermore, when we focus on UK universities, as we do in the later part of this chapter, we avoid both false positives and false negatives as the academic domain (.ac.uk) is stable and predictable in a way that the commercial domains are not. Essentially, all universities in the United Kingdom have a main address in the .ac.uk domain, and almost all addresses in the .ac.uk domain are universities (with a few exceptions for academic-affiliated organizations that are not themselves universities).[2]

Another issue that must be decided when undertaking analysis of web domains is the appropriate level of detail. This includes the temporal resolution to use for analysis (since while the web is constantly changing, the number of snapshots available in Internet Archive data vary over time based on the crawl settings in place when the data were gathered). In addition, the level of detail to be extracted from web pages must be determined (i.e. the appropriate level of resolution of page content, link information, page metadata, and so forth). Previous research on the .uk ccTLD has examined monthly snapshots over a one year period, finding that page-level hyperlinks change frequently month to month (Bordino et al., 2008). As Brügger (2013) notes, there are several reasons why archived websites are different from other archived material in respect to these details: choices must be made not just about what to capture but there are also technical issues about what can be archived and how the archiving process itself shapes the later availability of the archived materials.

Previous research using national web archives

While there have been a number of papers describing the practices of constructing national web archives (see for instance Masanès, 2005; Gomes et al., 2006; Baeza-Yates et al., 2007; Žabička and Matjka, 2007; Aubry, 2010; Hockx-Yu, 2011; Rogers et al., 2013), there are few that report using national web archives using large-scale (or even medium-scale) computational methods.

Thelwall and Vaughan (2004) used data from the Internet Archive to assess international bias in the coverage of the archive's collection. At the time of their study, however, it was not possible to access the data in the archive via automated means, so they were limited to relatively small samples of between 94 and 143 websites for each of four countries (total $N = 382$), accessed via the public Wayback Machine interface. They determined with these methods that there was an unbalanced representation of different countries in the archive, partially explained by technical factors rather than by biased policies.

The *Analytical Access to the Domain Dark Archive* (AADDA) project[3] and then later the *Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research* project[4] and the *Big UK Domain Data for the Arts and Humanities* project[5] enabled researchers to use UK Web Archive data for analytical study. These projects also demonstrate one of the legal issues of working with web archive data: the UK web archive data held by the British Library can be made available to researchers for use, but full-text content is only available via systems at the British Library. The raw data in the ARC/WARC files cannot be moved outside the Library's computer systems. As a result, many of the demonstrator projects that came out of these bigger projects focused on more qualitative, close analysis (see for instance Gorsky, 2015; Huc-Hepher, 2015) that was *enabled* by computational methods involving search, indexing and ontologies created by the project developers, the actual researchers largely used the extracted results in non-computational ways (see Chapter 11). It is important to note, however, that derivative datasets such as the list of web pages in the archive and the list of hyperlinks can be distributed more widely, which enables some large-scale approaches as we do in this chapter.

Another European project on *Longitudinal Analytics of Web Archive Data*[6] published a number of technical reports and papers that demonstrate computational approaches to working with web archive data but, as far as we are able to determine, there have not been the same sort of domain investigations as those done using the tools we report here.

The lack of studies using web archives in general, and using large-scale computational approaches in particular, has been documented in earlier work by members of this team (Dougherty et al., 2010; Thomas et al., 2010; Meyer et al., 2011; Dougherty and Meyer, 2014). In those papers and reports, we found that there remains a disconnect between the relatively active community engaged in archiving the web, and the relative lack of any community forming around large-scale analysis of web archives. This study is in part an attempt to fill that very clear gap.

The UK web domain

The .uk country-code top-level domain is managed by the internet registrar Nominet.[7] Below the .uk top-level domain are several second-level domains (SLDs), the largest of which are .co.uk (commercial enterprises), .org.uk (non-commercial organizations), .gov.uk (government bodies), and .ac.uk (academic establishments).[8] This chapter examines third-level domain data such as nominet.org.uk (Nominet), fco.gov.uk (the Foreign and Commonwealth Office of the UK government), or ox.ac.uk (the University of Oxford).

In the case of web archives (or indeed of other archived material which takes the approach of archiving all that can be archived, without a particular topic in mind), it is not scholarly interest in any particular topic that has set the data collection agenda. Instead it has been the goal of the archiving institution to accumulate material for the sake of preservation, leaving the question of the eventual uses of the archive data to later researchers. This means that the scope of the archived material and the level of detail available, as with other historical materials, is a function of the archiving processes used to gather and store the data. Thus, unlike web archive research done on the live web using researcher-implemented data collection mechanisms (e.g. Escher et al., 2006; Foot and Schneider, 2006), for the purpose of this study the dataset itself should be seen as a given. However, it can be mentioned that the Internet Archive's data comprise the most comprehensive archive of the web available (Ainsworth et al., 2011).

It is important to note that while the Internet Archive (IA) is the *most* comprehensive archive of the web available, that should not be confused with thinking that the IA crawls represent a *fully* comprehensive record of the web. The data collected over the 15-year period we are examining used a variety of methodologies and were done at varying levels of granularity. Data from the earliest years came from Alexa with 'no visibility into how this data is crawled', and the IA obeys robots.txt restrictions set by site owners (Jisc, n.d.-b), which can result in some websites missing pages or even being excluded completely from the archive (see chapter two by Hale et al.). The time between crawls is variable for any given page, resulting in some pages having more captures over time than others. Furthermore, the Internet Archive does not use the zone file from Nominet, which forms a complete list of all domains within .uk. Instead the Internet Archive relies on discovering websites through hyperlinks and other methods.

Data

Data preparation

The data for this study originally come from the Internet Archive, which began archiving pages from all domains in 1996 (Kahle, 1997). For the .uk domain that will be examined here, the data are sourced from copies of the approximately 30 terabytes of compressed archive data relating to the UK domain (the .uk ccTLD). Archive files were provided to the British Library by the Internet Archive with the specific purpose of creating the basis of a national archive of the web in the UK. These data form the 'Jisc UK Web Domain Dataset' (Jisc, n.d.-a).[9] The data provided to the research team by the British Library do not include the full text of all the pages crawled due to legal restrictions on use outside the British Library, but do include the link data and other metadata extracted from the full archive.[10]

The data were cleaned by removing error pages (e.g. 404 Not Found pages) as well as pages not within the .uk ccTLD. This resulted in a plain-text list of all page Uniform Resource Locators (URLs) remaining in the collection and the date and times they were crawled, and an additional plain-text list of all outgoing hyperlinks starting from pages within the dataset.

For this study, we started with this list of hyperlinks and filtered it to only include links between different third-level domains. We further grouped pages crawled at similar times (within 1,000 seconds) together and assigned the hyperlink pair a weight based on the number of hyperlinks between the two third-level domains in that time period. For each year, if there are multiple crawls within the dataset we take the crawl with the largest number of captured hyperlinks between any two domains. We also formed one list of all third-level domains present in the dataset each year and the number of pages crawled within each third-level domain. These data were loaded into Apache Hive for the analysis that we present here.

Data analysis

In what follows, we undertake a longitudinal network analysis, charting the .uk domain and its core second-level domains over time. As Brügger (2013) points out, this type of analysis is not concerned with who produced what, nor with how the web content was used, but rather with what was created and thus 'the web which is' – or was – 'actually available to users'.

First, we present an overall longitudinal view of the second-level domains within the .uk domain. We investigate the growth of the entire domain between 1996 and 2010, broken down into its four largest constituent parts, .co.uk, .org.uk, .gov.uk, and .ac.uk. Analysis of these SLDs allows us to investigate the role of different sectors of UK society in the growth of the UK web presence.

The second section looks at the link density within and between second-level domains. We examine the internal link density of each SLD, and analyse how they interact with each other: whether, for example, there are more links between certain subdomains, and whether linking is reciprocal between domains or whether it is unbalanced.

The third and final section of the findings takes a closer look at the academic second-level domain .ac.uk. This research builds on earlier longitudinal analyses of academic web pages, which have investigated, for example, the stability of outlinks (Thelwall et al., 2003; Payne and Thelwall, 2007). Our findings update earlier studies by extending the period of analysis to the end of 2010 and assessing the effect of new variables, including institutional affiliation, league table ranking and geographic location on link practices between different universities.

Results

Overview of growth in the .uk web domain

Figure 1.1 displays the overall growth of the .uk ccTLD, showing the total number of nodes (on a logarithmic scale) within each of the four main SLDs we analysed over the period from 1996 to 2010. The insert in the figure shows the size of the entire .uk domain (on a linear scale). There is a clear change in the trend of the growth around 2001 for .co.uk and .org.uk as both domains continue to increase in size, but at a lower speed. Furthermore, .ac.uk and .gov.uk seem to almost stabilize in size at around the same time.

Figure 1.2 shows the relative size of the second-level domains .co.uk, .org.uk, .ac.uk, and .gov.uk across the 15-year period, standardized as each SLD's proportion of the total nodes (i.e. domains/websites, not web pages) in the collection in each year. While these are not the only second-level domains in use within the .uk domain, they are the four largest in terms of number of nodes across the whole period.

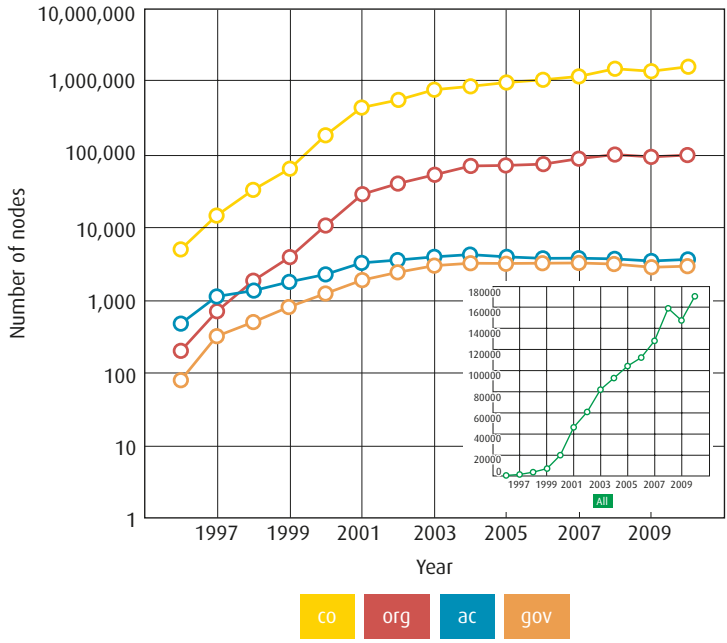

Figure 1.1 Number of nodes (third-level domains) within each second-level domain over time. The inset shows the sum over all second-level domains

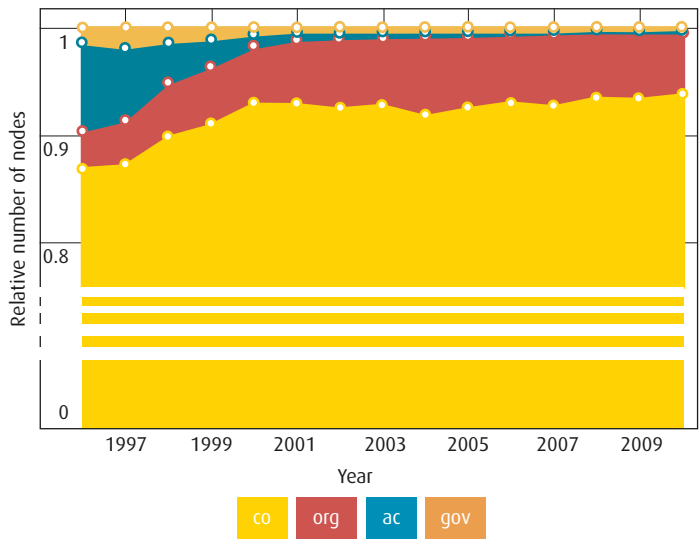

Figure 1.2 Relative size of second-level domains in the .uk top-level domain over time

As Figure 1.2 shows, .co.uk is the predominant second-level domain throughout the entire period, with .co.uk sites never accounting for less than 85% of the total. However, also apparent is the large proportion of governmental and, especially, academic sites in the early recorded history of the UK web. This is consistent with the role that universities played in the early establishment, adoption and development of the web (Leiner et al., 2009). Over time, however, this early presence was greatly overshadowed in terms of absolute numbers of nodes when compared to the continued growth of the .co.uk and .org.uk domains.

Link density within and between second-level domains

Up to this point the analysis has drawn only on node data; that is, the number of websites making up each domain. However, link analysis can offer insight into how well connected each SLD is with itself and with other domains. A link from one site to another has been used as an indicator of awareness between blogs (Hale, 2012) and recognition between academic sites (Thelwall et al., 2003). Figure 1.3 shows, for each subdomain, how many total links there are for every node over time, where a fluctuating relationship between the number of nodes and links to other nodes for each second-level domain is visible. Over the whole period, the .ac.uk academic SLD and, from 1997 onwards, the .gov.uk governmental SLD are the most internally dense SLDs. This observation may reflect the fact that registration for the .ac.uk and .gov.uk subdomains is restricted, whereas .org.uk and .co.uk sites can be registered easily by any party. In addition, the .ac.uk and .gov.uk subdomains are likely constituted by a narrower and more cohesive set of institutions, creating, on average, a stronger basis for linking within the SLDs. Furthermore, there is likely more competition and thus less reason to link within the .co.uk commercial subdomain compared to .ac.uk or .gov.uk. Higher link density within the .org and .gov domains in comparison to the .com domain has previously been observed during a smaller scale, topical study about climate change (Rogers and Marres, 2000).

Also of note is the general rise of links in the middle of the period, particularly in the substantial .co.uk subdomain. This peaks sharply in 2004 before falling sharply back to around pre-2001 levels by 2009. This trend has no easy explanation, suggesting that further research is required to explain this pattern. Possible explanations include that the norm of including lists of links on web pages such as blogs fell out of favour in the middle of this period or that more websites increasingly linked outside of the .uk ccTLD.

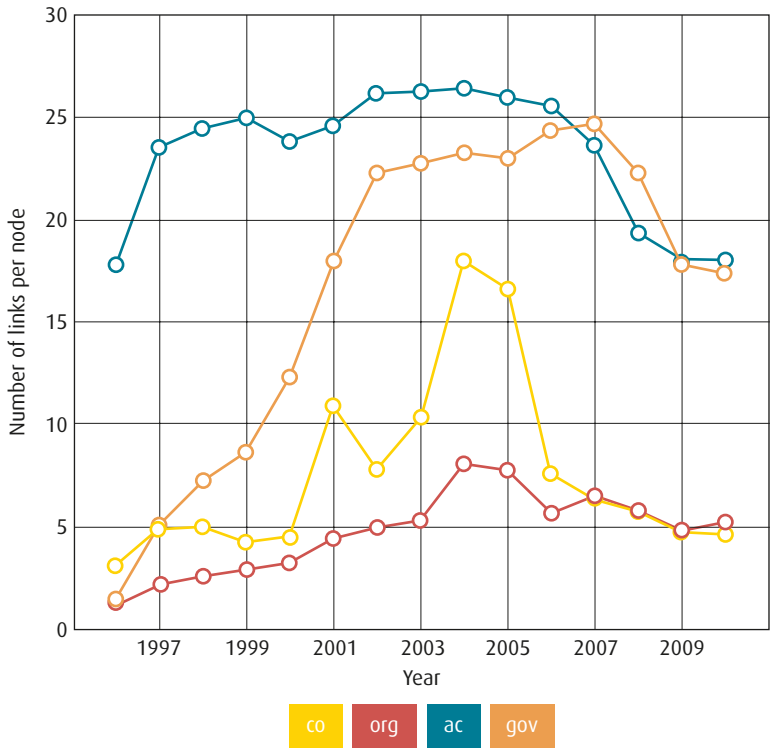

Figure 1.3 Number of within-SLD links per node in four .uk SLDs, 1996–2010

Not only can web domain data tell us how well integrated an SLD is internally, but we can also investigate how well SLDs are connected to each other. Figures 1.4a and 1.4b show the quantity of links between SLDs for 2010, the last year in the dataset, where the size of an arc relates to the volume of links from one SLD to another. The colour of each arc relates to links sent in one direction, from the host SLD outwards. For example, green arcs show links from the .co.uk domain to others. Figure 1.4a shows the absolute volume of links, while the size of the arcs in Figure 1.4b are normalized in relation to the number of nodes in the target subdomain. (Note that Figure 1.4a does not display links within a single SLD, as the volume of links between .co.uk sites dwarfs all other relationships. As Figure 1.4b controls for the number of nodes in each SLD, the adjusted .co.uk arc is much smaller and links within a single SLD are therefore included.)

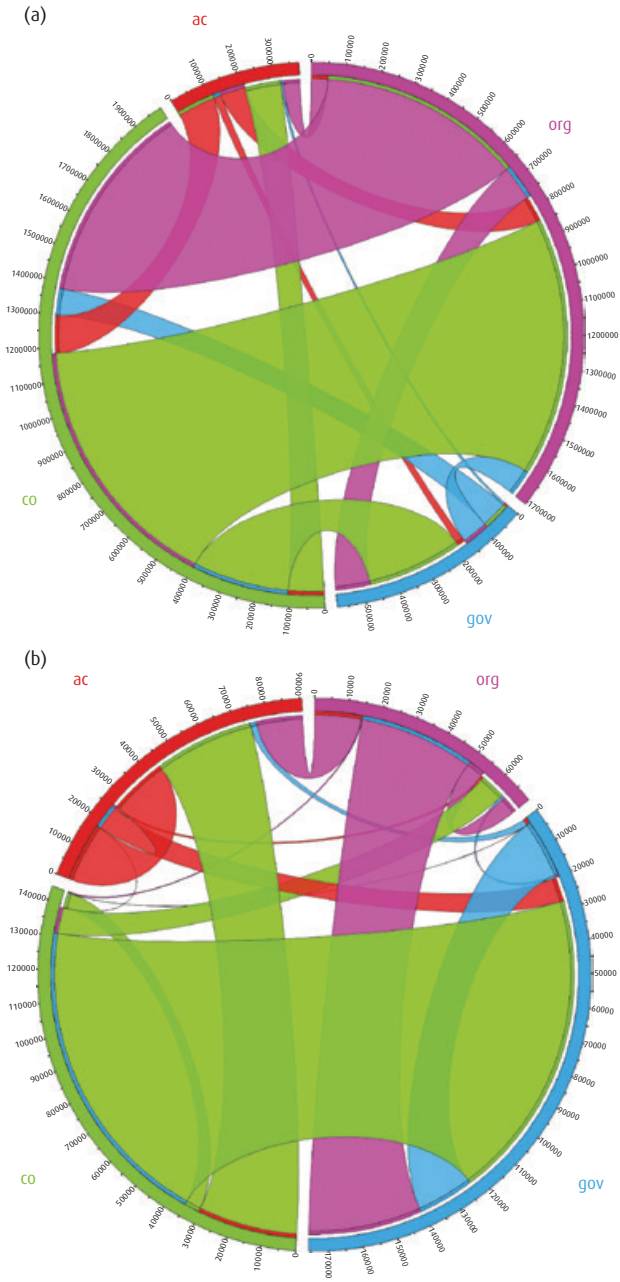

Figure 1.4 Links between four second-level domains. Panel *a* shows the absolute number of links between different SLDs (self-loops are excluded), and panel *b* shows the relative number of links normalized by the size of target subdomain

Figure 1.4a shows that the largest volume of links between SLDs in 2010 flowed from .co.uk sites to .org.uk sites, and this relationship is fairly reciprocal, with .org.uk sites sending almost as many links back. Links between other domains are much lower in terms of absolute volume. When controlling for the size of the target subdomain, however, the picture changes somewhat. As Figure 1.2 showed, by 2010 the number of nodes in the .org.uk subdomain far outweighed those in the .ac.uk and .gov.uk subdomains. Figure 1.4b, adjusting for this, shows that the .gov.uk and, to a lesser extent, the .ac.uk subdomains punch above their weight, receiving proportionally more links from .co.uk and .org.uk sites. Once again, the more restrictive registration policies for these SLDs may be a factor here, driving up the average quality and 'linkworthiness' of sites in these subdomains compared to .co.uk and .org.uk sites. However, this discrepancy may also be related to other factors such as the comparative homogeneity of these SLDs, the perception of objectivity or balance on academic or government websites as opposed to sites oriented towards sales or persuasion, or even the international standing of many UK universities, although understanding these factors would require further investigation.

For the .gov.uk subdomain, the finding that sites link out less than they are linked to suggests a lack of 'outward-lookingness', compared to the other sectors. In contrast, Escher et al. (2006) found the UK Foreign and Commonwealth Office to be relatively more outward-looking than its equivalents in Australia and the USA. However, foreign offices, given their outward facing role, could easily be an exception to a more general government-wide propensity not to link out.

In addition, it is worth noting the relatively heavy proportion of links within the .ac.uk SLD shown in Figure 1.4b in the red arc that curves from 'ac' back into 'ac'. This propensity of academic institutions to link heavily to other academic institutions (more so than the other domains) reflects (taking a positive view) a strong network among academic institutions, but also potentially (taking a negative view) a tendency towards inward-looking, within-domain links. We examine these links in more depth in the next section.

The UK academic subdomain

At this stage we turn our attention to one particular subdomain, the .ac.uk academic subdomain of the UK web. To be eligible for a third-level domain within .ac.uk, an organization must have a permanent physical presence in the UK and either have the majority of its activities publicly funded by

UK government funding bodies or be a Learned Society. In addition, the organization must satisfy at least one of the following criteria: the organization must provide tertiary-level education with central government funding, conduct publicly funded academic research, have a primary purpose of supporting tertiary-level educational establishments, or have the status of a Learned Society ('a society that exists to promote an academic discipline or group of disciplines').[11]

The academy was at the forefront of the development of the web, and, as Figure 1.2 shows, .ac.uk sites constituted a sizeable minority of .uk sites in 1996. Over time, this proportion waned, even as more UK universities established a substantial web presence. In this subsection we use the longitudinal data collected to examine the relationship between universities' linking practices and three variables: institutional affiliation, league table ranking and geographic location. Our hypothesis in doing so was that higher status academic institutions would be more strongly linked to than lower status institutions and would also be more strongly interconnected with their peer institutions.

For the analysis, we built a list of the 121 universities listed in the 2014 *Sunday Times* University Guide.[12] Each of these universities has a website, all of which use the .ac.uk suffix. We obtained the third-level domain (e.g. ox.ac.uk) for each. Further data collection as necessary is described in the respective subsections that follow.

Group affiliation

Many UK universities belong to associations, formed to represent their interests and facilitate collaboration. The groups are neither mutually exclusive nor exhaustive, meaning that universities can belong to none, one or more than one group, but for practical and political reasons most universities belong to only one. We collected data on the memberships of five groups, the Russell Group,[13] the 1994 Group,[14] the University Alliance,[15] the Million+ Group,[16] and the Cathedrals Group.[17]

The best known of these is perhaps the Russell Group of research-intensive, highly ranked universities, formed in 1994 and now constituted of 24 members. The 1994 Group, which represented smaller research institutions, was formed in response to the Russell Group, but disbanded in 2013. Given the time frame of the dataset we include the 11 final members of the group in our analysis. Of the remaining three groups, the University Alliance is formed of 22 business-oriented UK universities, the Million+ Group is made up of 17 mostly 'new' (post-1992) institutions, and the Cathedrals Group is made up

of 16 universities originally instituted as church-led teacher training colleges. The stated purposes of these groups differ somewhat, but each are constituted broadly to serve the research and educational interests of their members.

In comparing group membership to the density of links between different universities, we sought to discover whether academic affiliation was associated with the density of links between institutions. To do this, we performed a network analysis, investigating whether the universities clustered on the basis of group affiliation. Figure 1.5 shows a network diagram, with different affiliations marked by different colours.

To the naked eye, Figure 1.5 shows no discernible clustering on the basis of group affiliation, and network analysis bears this out. The division of the network by affiliations has a modularity score (Newman, 2006) of −0.003, indicating that the division of the network into clusters based on university affiliation is no better than dividing the network into five random clusters. On an individual basis, only one group, the Russell Group, has many internal links and comparatively fewer links to institutions outside the group. It is the most strongly connected group with an internal hyperlink density of 0.71. The Russell Group, which includes 24 of the leading international UK universities with some of the highest levels of research funding, arguably represents most if not all of the elite universities in the UK. It contains nine of the ten top-ranked UK universities, including both Oxford and Cambridge. That these universities are more strongly linked to each other is likely related at least in part to their active research cultures, with many collaborations existing between researchers at these top institutions. The lack of strong web connections in the other associations, however, suggests that while these institutions may or may not have strong connections among their members by other measures, there is no evidence that universities strongly link to the websites of institutions with which they share group affiliation over institutions outside of the group.

League table ranking

University league tables are an important if imperfect indicator of a university's prominence. Modern league tables incorporate a whole range of measures, including factors related to teaching, research and student satisfaction. As such, we investigated whether a university's league table ranking is associated with its web presence, and whether the relationship has changed over time, in terms of both increasing adoption and development of an institution's web presence and its changes in league

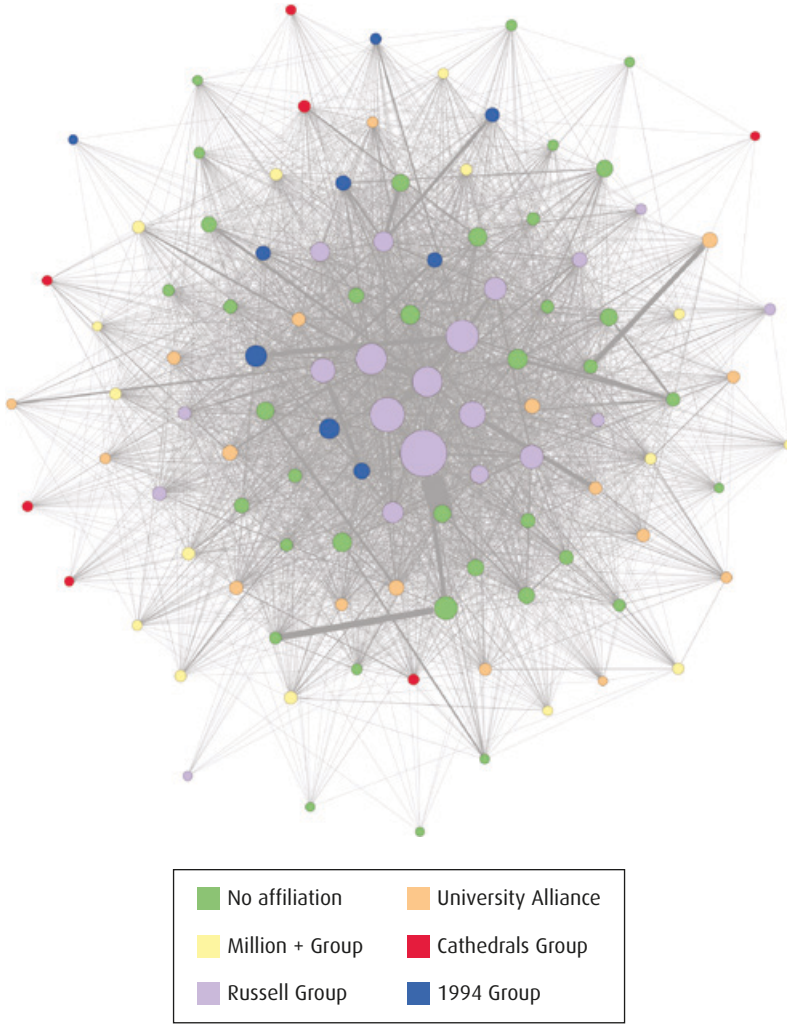

Figure 1.5 Network diagram of hyperlinks between universities. Different colours indicate different university affiliations

table ranking over time. For this analysis, we collected the rankings of UK universities published in *The Times* Good University Guide for three years, 2000, 2005 and 2010, and compared these rankings with data from crawls conducted in the same three years.

In conducting the analysis, we used ten common measures of network centrality for each of the three different years to gauge the relationship between each university's league ranking and its position in the

network of hyperlinks flowing between university third-level domains. We then produced lists ranking the universities for each year by each centrality measure and computed Spearman's rank correlation coefficient for each centrality ranking and league table ranking combination. These correlation coefficients are shown in Figure 1.6.

For most measures of centrality used, a pattern emerges: the data for 2010 show the strongest correlation between league table ranking and centrality, while the relationship is less evident for 2000 and 2005. The most strongly correlated measure is in-strength, a sum of all the hyperlinks linking to a given web domain. This measure uses the weight of each edge, which corresponds to the number of hyperlinks between any two third-level domains. This differs from in-degree which measures the number of other domains that link to a given web domain. Figure 1.7 shows the fairly strong correlation between universities' league table rankings and their network positions as measured by in-strength. What Figure 1.6 and Figure 1.7 suggest is two-fold: first, that a university's prominence, as measured by its league table position, is an increasingly stronger predictor of the number of links to that institution over the 2000–2010 period. Whether this is an example of the Matthew Effect ('the rich get richer') (Merton, 1968) whereby highly prominent institutions become well-linked institutions largely as a result of their prominence (and conversely, marginal institutions become more marginalized as a result of their lack of prominence), or whether there is another independent factor at play here cannot be determined from these data. However, the second conclusion is clear: the hyperlink patterns within the UK academic subdomain support the notion that the web does not inherently challenge existing power structures. Instead, the saturation of the .ac.uk subdomain, in terms of the presence of essentially all possible academic institutions by 2003 (as shown in Figure 1.1), resulted in a subdomain in which network centrality closely mirrors prominence as measured by league tables by 2010.

Role of geography

Finally, we investigated whether any association exists between the geographic proximity of UK universities and the density of hyperlinks between them. This analysis builds upon work by Pan et al. (2012) who found, at a global scale, that rates of academic citations and collaborations between two cities diminish as the distance between them increases, following gravity laws. We conduct a similar analysis,

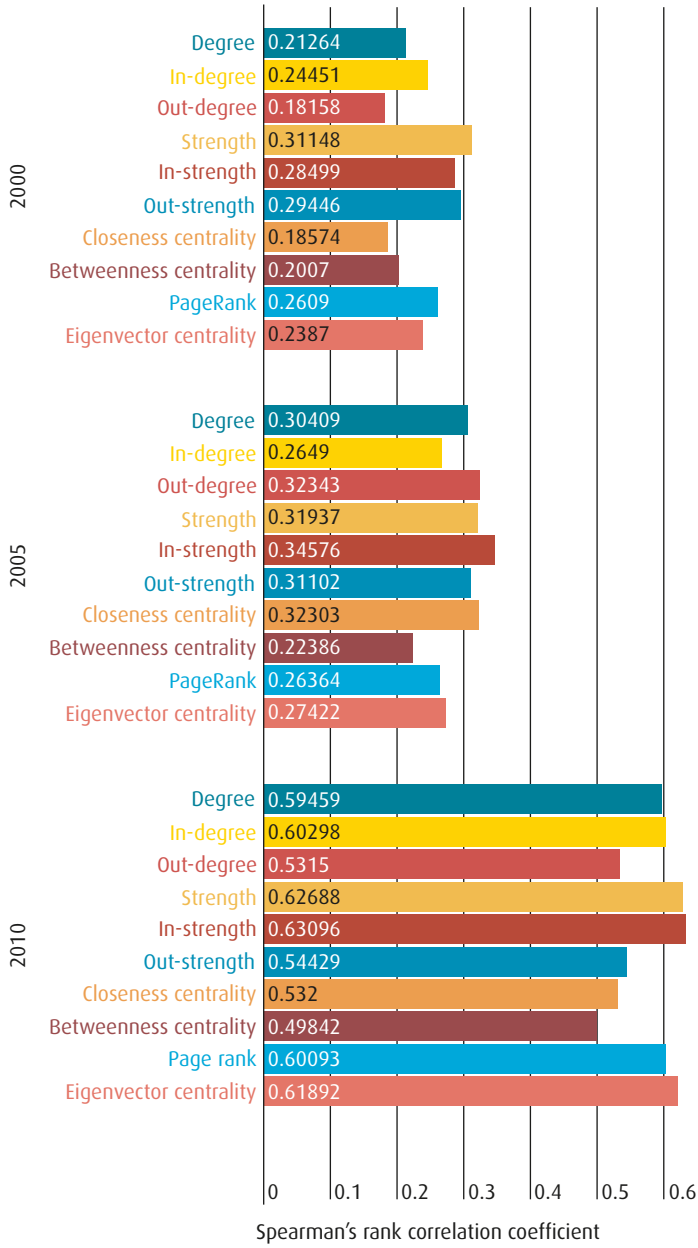

Figure 1.6 Spearman's rank correlation coefficients between university league table rankings and ten different network centrality measures for three years

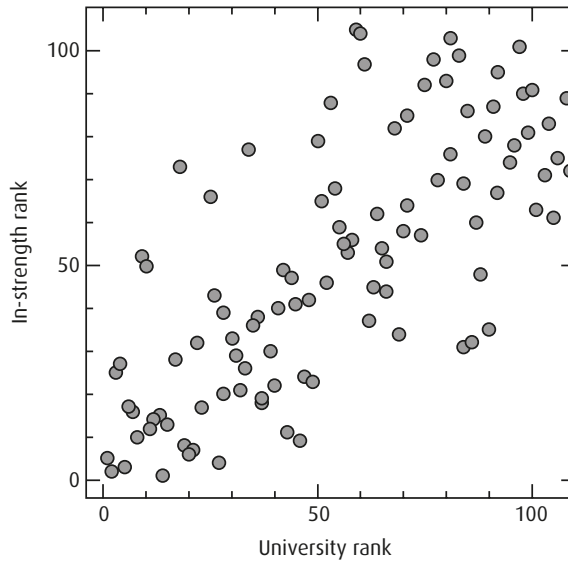

Figure 1.7 University in-strength rankings compared to university league table rankings for 2010. Spearman's rank correlation is 0.63

replacing citations and collaborations with hyperlinks collected in the web domain data.

We collected geographic coordinates for the UK universities in the list using simple Google Maps searches. Universities can be spatially complex, sometimes having multiple campuses and satellite sites; so, some discretion was occasionally required in identifying the centre of each university.

The standard, naïve gravity law approach would suggest that the number of hyperlinks, or the strength of the connection, between two given universities is inversely proportional to the square of the distance between the two universities. We let S_{ij} denote the strength from university i to university j. Focusing on the data from 2010, the left frame of Figure 1.8 shows that the relationship between this measure and the geographical distance between the two universities is very noisy. To correct for the different sizes of universities and their different linking practices (some universities may just link more than others), we normalize these strengths. We divide S_{ij} by the sum of the weights of all edges coming from university i (S_i^{out}) multiplied by the sum of the weights of all edges linking to university j (S_j^{in}). We denote this normalized measure σ_{ij} and plot it against physical distance in the right frame of Figure 1.8. With this normalization, the relationship between distance and the

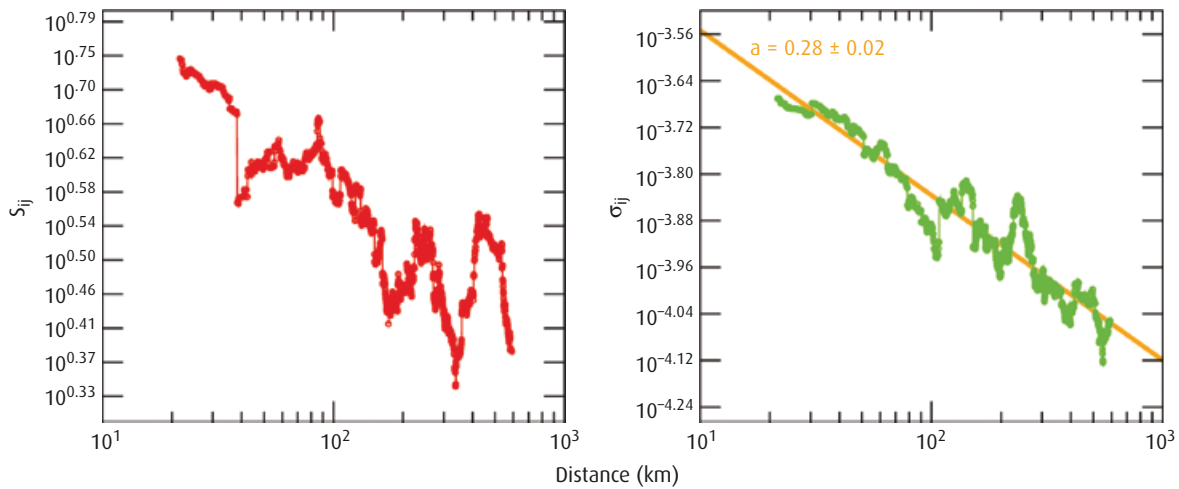

Figure 1.8 Left: Raw hyperlink strength (S_{ij}) between universities versus geographical distance. Right: Normalized hyperlink strength (σ_{ij}) between universities *versus* geographical distance. The normalized measure follows a gravity-law model with an exponent of a=0.28±0.02

number of hyperlinks (strength) between universities is very clear. In both frames, we use a moving average window with a length of 500 data points and therefore a lower bound of 20km is introduced. An upper bound is induced by considering only the universities within the UK in this study. However, the gravity law holds significantly within a large distance range of 30–600km.

Letting d_{ij} denote the geographical distance between two universities, we then seek the exponent a, which best fits the observed data following $\sigma_{ij} \propto d_{ij}^{-a}$. Using the least squares method, we fit a linear function to the logarithmically transformed data and find $a = 0.28 \pm 0.02$, which closely matches the findings of Pan et al. (2012) for citation and collaboration networks. In that study, Pan et al. found an exponent of $a = 0.30$ for the citation network before any normalization, while finding an even stronger role for geographical distance ($a = 0.77$) after applying a similar normalization to the one we apply here.

Figure 1.9 maps the universities in the sample along with the connections between them coloured according to σ. It is evident, especially in the map of 2010, that the longer connections generally have weaker strength. It is worth nothing that the size limit of the dataset and the geographical constraints—such as the dense region of London extended to Oxford and Cambridge, which includes a large number of universities in our dataset – could partially drive the strong geographical dependency we observed. This dense region is particularly visible in the map of 2005 in Figure 1.9.

Conclusion

In this chapter we have reported findings based on longitudinal analysis of the recorded history of the UK web domain from 1996 to 2010. While this analysis is by necessity at a macro-level in terms of detail, it nevertheless demonstrates the potential of these data for detecting changes in patterns in web linking behaviour over time. Such evidence is related to the growth and expansion of the web and uneven patterns of linking within subdomains, such as the academic .ac.uk subdomain discussed in this chapter. We have shown that even though the growth of the commercial side of the web has resulted in increasing commercial dominance of the UK ccTLD in terms of absolute number of nodes, the academic and government subdomains receive proportionally more inlinks per domain. In examining the academic subdomain in particular, we have shown that while there is no generalized

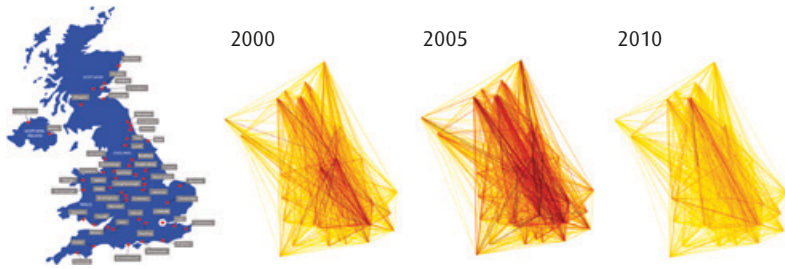

Figure 1.9 Maps of the UK universities under study for three years: 2000, 2005 and 2010. The connections are the hyperlinks and colour corresponds to the normalized strength of each link (σ_{ij}). The reddest links correspond to the strongest connections

clustering based on the affiliation of academic institutions, there are clear patterns in terms of a higher number of inlinks to academic institutions with higher statuses and stronger connections between geographically-closer institutions.

This research has also demonstrated some of the benefits and challenges of this type of analysis. The methods and results described here have allowed us to paint a reliable portrait of the .uk web domain over a period of growth spanning 15 years, which would otherwise be impossible without using web archives (unless a researcher had started collecting similar data themselves over the same time period, which could work going forward, but not retrospectively). We have also shown that it is possible, within the limits of an admittedly incomplete national web archive, to understand certain domains in greater detail, as we have done with the academic portion of the UK web domain.

Challenges, however, remain. Working with these data was neither simple nor quick, and the link data required significant cleaning before they were usable. Also, while the file structure for the link data was very simple, the sheer size of the data necessitated the use of larger processing infrastructure (Apache Hive) that not all researchers have access to or the skills to use. Further, because of legal limitations on the distribution of actual page content, questions that arose over inconsistencies in the link data that might have been easier to understand by looking at the context of the link were more difficult to resolve.

The biggest challenge, however, to using web archives in computational ways remains finding the right questions that are both interesting and capable of being answered within the limits of the web archive data and the extent to which any given web archive contains appropriate coverage over the time period of interest.

This analysis suggests many future possibilities for research with these web archive data, including more detailed micro-level analysis of linking behaviour within various subdomains over time, discovery of networks of collaboration between subunits of institutions, comparison between link measures and other measures of prominence such as citation networks and analysis of other subdomains besides .ac.uk. In addition, there are ongoing efforts to prepare the full-text corpus extracted from the web archive for research (rather than the link corpus used here), which it will be possible to combine with these data to answer more detailed questions about the content of the web, the context for links and discourses on the web.

Acknowledgements

The authors would like to thank Ning Wang for his advice and support on data cleaning for the original project and Andreas Kaltenbrunner for his help with creating the original geographic visualizations. The authors are also grateful for funding from UK Jisc for the 'Big Data: Demonstrating the Value of the UKWeb Domain Dataset for Social Science Research' grant (16/11 Enhancing the Sustainability of Digital Collections) that supported the data extraction and early analysis, and further funding for analysis from the UK Arts and Humanities Research Council for the 'Big UK Domain Data for the Arts and Humanities' grant (AH/L009854/1). Finally, the authors would like to thank our anonymous reviewers for their helpful comments on both this chapter and the earlier version of this research, which was published in the Proceedings of the 2014 ACM Conference on Web Science (Hale et al., 2014) and is updated here with permission from ACM.