The Web Structure of E-Government - Developing a Methodology for Quantitative Evaluation

Vaclav Petricek
UCL Computer Science
Gower Street
WC1E 6BT London
+44 20 7679 0418
v.petricek@cs.ucl.ac.uk

Tobias Escher
UCL School of Public Policy
29/30 Tavistock Square
WC1H 9QU London
+44 20 7679 4903
t.escher@ucl.ac.uk

Ingemar J. Cox
UCL Computer Science
Gower Street
WC1E 6BT London
+44 20 7679 7608
ingemar@ieee.org

Helen Margetts
Oxford Internet Institute
1 St. Giles
OX1 3JS Oxford
+44 1865 287207
helen.margetts@oii.ox.ac.uk

ABSTRACT

In this paper we describe preliminary work that examines whether statistical properties of the structure of websites can be an informative measure of their quality. We aim to develop a new method for evaluating e-government. E-government websites are evaluated regularly by consulting companies, international organizations and academic researchers using a variety of subjective measures. We aim to improve on these evaluations using a range of techniques from webmetric and social network analysis. To pilot our methodology, we examine the structure of government audit office sites in Canada, the USA, the UK, New Zealand and the Czech Republic.

We report experimental values for a variety of characteristics, including the connected components, the average distance between nodes, the distribution of paths lengths, and the indegree and outdegree. These measures are expected to correlate with (i) the navigability of a website and (ii) with its "nodality" which is a combination of hubness and authority. Comparison of websites based on these characteristics raised a number of issues, related to the proportion of non-hyperlinked content (e.g. pdf and doc files) within a site, and both the very significant differences in the size of the websites and their respective national populations. Methods to account for these issues are proposed and discussed.

There appears to be some correlation between the values measured and the league tables reported in the literature. However, this multi-dimensional analysis provides a richer source of evaluative techniques than previous work. Our analysis indicates that the US and Canada provide better navigability, much better than the UK; however, the UK site is shown to have the strongest "nodality" on the Web.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services
H.5.4 [**Hypertext/Hypermedia**]: Navigation, Architectures
K.4.1 [**Public Policy Issues**]

General Terms

Measurement, Performance, Design

Keywords

e-government, national audit offices, ranking, webmetric, network analysis, quantitative evaluation

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

1. INTRODUCTION

Relationships between consumers and commercial organizations of all kinds have been revolutionized by the phenomenal rise of e-commerce. Similarly, the increasingly widespread use of the internet and the World Wide Web offers a potential transformation of government-citizen relationships in the development of 'e-government' – the use by government of information technologies both internally and to interact with citizens, businesses and other governments.

Most advanced industrial nations have put considerable political support and financial resources behind the development of e-government. By 2005, the UK for example has a '.gov' domain of around 8 to 23 million pages (depending on which search engine estimates one tends to believe, MSN or Google respectively) and spends £14.5 billion a year on information technology in the pursuit of the Prime Minister's commitment to have all government services electronically available by the end of 2005.

In spite of these resources (greater than 1 per cent of GDP in most industrialized nations is spent on government information technology), e-government has lagged behind e-commerce. In the UK, recent survey evidence [8] suggests that while 85% of Internet users claim to have looked for or bought goods and services online, and 50% of users to shop online at least once a month, only 39% have had any sort of interaction with government online in the last year. While figures for e-government usage are much higher in some countries, particularly Scandinavia, the generalization that government has been far less touched than commerce by widespread use of the World Wide Web holds true internationally. Governments are under pressure to demonstrate that the massive investments they are making are worthwhile.

Furthermore, lack of progress in e-government can affect a government's policy-making capacity. One of the key 'tools' of public policy deployed by government has been defined within the field of political science as 'nodality' – the characteristic of being at the centre of social and informational networks [11][12]. The concept of 'nodality' in political science is analogous to authoritativeness (often indicated by number of links pointing to a site) and hubness (number of links pointing outside a site) with respect to computer science and the Web. Intuitively, we would expect government to become more nodal as the Internet and associated technologies become more embedded into all aspects of social and political life. However, if private sector organizations and non-governmental organizations are more successful at using the World Wide Web to increase their nodality, it may be that government will suffer a net loss of

nodality in the virtual realm, thereby weakening one of its key tools. We can hypothesize that a 'healthy' government domain, if we can establish appropriate characteristics to define such a thing, will help government to become more nodal. If a domain has more incoming links, for example, it is likely to be more visible to search engines and more easily found by citizens searching for government-related information.

Current research in computer science, political science or communications research tells us little about the Web structure of e-government. As outlined below, there are numerous qualitative analyses of government online and international rankings of e-government. However, there are no quantitative attempts to analyze the underlying link structure of a government domain, to assess the 'health' of such a domain, or to compare domains. We believe that such an undertaking will provide valuable information about this new element of government and ultimately could be used to aid the redesign of governments' online presence. Although there are a number of studies developing the field of webmetrics (discussed below), there have been virtually no attempts to apply them to government.

Existing Web metrics may be categorized as either user-based or structural. User-based metrics measure a variety of characteristics of a user's usage of a site, e.g. page impressions. Acquisition of user-based metrics can be difficult since the data is only available to the owner of the website, and the data may be difficult and costly to obtain. Furthermore, cooperation might not be forthcoming, if the evaluation is performed by a neutral third party intent of publishing comparative data between sites. And data may be unavailable for reasons of user privacy or contract confidentiality (where external providers maintain the site).

In contrast, structural metrics measure properties such as the average distance between two random pages and the interconnectedness of sites. They are readily available to anyone capable of crawling websites. They would seem to offer the potential for establishing the 'health' of a domain. For example, if a government domain is highly inter-connected, then citizens are much more likely to find information (such as how to make an application for a visa) by traversing the link structure of the site.

There are numerous technical and methodological problems involved in creating such metrics at the domain level. The numbers of pages, nodes and links involved can be very large. For this reason, we have chosen to test our methodology on a single agency example – the national audit office – which may be compared across several countries. The UK National Audit Office, the US General Accounting Office, the Canada Audit Office, the Controller and Auditor General of New Zealand and the Supreme Audit Office of the Czech Republic all have roughly comparable roles and responsibilities. Thus, we believe that performing a comparative evaluation of their online presence will take us some way towards developing a methodology for the much larger task of comparing whole government domains. In addition, the audit offices can be deemed to constitute some part of each country's overall e-government effort, so there will be some value in comparing existing evaluations of each country's place in e-government rankings.

In Section 2 we briefly survey prior work in the field. Section 3 then discusses our experimental methodology. Section 4 describes experimental results and Section 5 provides a discussion.

2. PREVIOUS WORK

There have been extensive efforts to assess the quality of e-government throughout the world. An overview of this work is provided in Section 2.1. Similarly, there have been numerous studies within the computer science community to assess and characterize the structure of hyperlinked environments, as discussed in Section 2.2.

2.1 Qualitative Assessment of E-Government

There have been numerous attempts to assess e-government internationally, in the form of rankings of countries carried out or commissioned by international organisations (such as UNPAN [15], European Commission [4]), private sector consultancies (particularly by Accenture [1] Taylor Nelson Sofres [14] and Graafland-Essers and Ettedgui [9]), and academic commentators [8][16][13][6]). While some are widely cited and eagerly awaited by governments which score well, all are of methodological questionability and rely, ultimately, on subjective judgments. Most make some form of assessment of government websites according to content (eg. [16]) and availability of services (eg. [4]), while Accenture's widely known annual study is largely a qualitative analysis based on researcher assessments of websites and available e-services and a limited number of short visits to the 22 countries covered.

All these studies fail to collect either user-based or structural Web metrics. None have been able to overcome the 'cooperation' problem, noted above, with regard to collecting user data for significant numbers of websites, although some use survey evidence (Taylor Nelson Sofres [14] in particular, while Accenture included a user opinion survey for the first time in 2005 [1]) to estimate the extent to which a population as a whole have interacted with their government online. West [16] gathered content-related data from approximately 2,000 websites in nearly 200 countries and LaPorte and Demchak developed measures of 'interactivity' and 'transparency' for tracking the diffusion and use of the Web in nearly 200 governments around the world (now discontinued). However, none of these studies have considered the link structure of e-government sites.

Methodological variations across these studies are evidenced by the different rankings that the countries achieve. The table below gives the scores attained by the countries covered here in the most recent reports by the UN [15], Accenture [1], Taylor Nelson Sofres (TNS) [14], and West [16]. The table shows that (for example) the UK scores well with the UN but close to the median for Accenture's 22 countries, well below Canada and the US for West and lowest of these six countries for TNS, the latter one reporting percentage of population using government online in the last 12 months.

Table 1. Selected results of important e-government rankings

	UN	Accenture	TNS	West
Maximum	***0.927***	***1st***	***63***	***46.3***
Canada	0.806	1st	51	42.4
Czech Republic	0.542	n/a	23	33.8
New Zealand	0.710	n/a	45	35.5
United Kingdom	0.814	10th	18	37.7
United States	0.927	2nd	44	45.3
Minimum	***0.009***	***22nd***	***1***	***16.0***
Countries Surveyed	***191***	***22***	***31***	***191***

However, if we triangulate the five countries across the sample of the three rankings given that include all countries in our sample (by giving 5 pts for each 1st place, 4 for 2nd etc. and adding them), we find that the US and Canada emerge in the top slot (with 13 and 12 respectively out of a possible 15), followed by the UK and New Zealand (with 9 and 8). The Czech Republic occupies the last place (3). Consequently, while we may not be able to discriminate between the USA and Canada, we are interested to determine whether structural metrics reveal distinctions between the US and Canada on one hand and the less well performing countries on the other.

2.2 Previous Web Metric Work

There have been many studies concerned with website usability from a user perspective. We do not review this literature here, as it is outside the scope of this paper, which is concerned only with the link structure of the sites and their neighborhood.

The idea of a link as an endorsement, inspired by bibliometrics, has been successfully applied to a wide range of problems from ranking algorithms [26][20][30][31], through focused crawling [2] to web page classification and clustering [30][32][33].

There have also been extensive studies investigating the structure of the Web [24][27][35], as well as proposals for its generative models [10][27][28][36], all of which noted the scale-free structure of the network.

Usually, the study of hyperlink structure has focused on academic networks [3][19]. Studies have benefited greatly from the methods developed for social network analysis (see for example [17]) and in recent years researchers from various areas have tried to apply these methods to the Internet by interpreting the relation between actors through the hyperlink connections of their websites [18].

The application of computer science methods to the study of politics on the web and e-government in particular is not yet very common, although there are some notable exceptions. For example, Hindman et al [10] studied the communities surrounding political sites and showed that (i) the number of incoming links is highly correlated with the number of actual users and (ii) that online communities are usually dominated by a few sites – winners who take all the attention. Overall, applications of computer science, and especially Web metrics, to the quantitative evaluation of e-government have not been reported.

3. METHODOLOGY

For purposes of clarity, Section 3.1 briefly defines the vocabulary used to describe the Web. Section 3.2 then describes the datasets used in our evaluation and how these datasets were acquired.

3.1 Definition of Terms

Networks have been studied in a variety of different fields, including computer science and social studies. The diversity of these disciplines has led to a diversity of vocabulary so a brief definition of terms seems useful.

A network consists of a set of *nodes* and a set of *directed links* that connect pairs of nodes. In our case, the nodes are documents retrieved from the Internet and the links are hyperlinks that can be used to navigate between these documents. Every link has a *link source* that is the node from which it originates, and a *link target*, which is the node to which it points. A node can be described in terms of its links: every node has an *indegree* which represents the number of links for which the node is a link target, and an *outdegree* which represents the number of links for which the node is the link source. The sum of the *indegree* and *outdegree* is the *degree* of a node. A node with non-zero indegree is receiving links or has *inlinks*, while a node with non-zero outdegree has *outlinks* or is pointing to another node.

We distinguish different entities on the Web that are ranked in a hierarchy. The smallest entity is a document on the Web, identified by a *Uniform Resource Locator (URL)*. A set of documents constitutes a *website*. A website is primarily conceptual and depends very much on the perception of the user or the author of the site. It is difficult to identify websites automatically.[1] In this paper, we will assume that a website is identified by a *host*, i.e. everything between http:// and the next slash stripped off generic 'www' prefix and any trailing port numbers. Although not perfect, this automated approach seems sufficient [21]. Hosts can be classified according to *domain level*. Reading from the right, every combination of letters preceded by a dot constitutes a level of domain.

3.2 Datasets

As noted above, the size of e-government sites can be very large. Furthermore, it is very difficult to define what e-government encompasses. For example, should the military, local government, broadcasting or health services be included? These issues have to be resolved before analysis at the level of a whole government can take place. For this preliminary study, we therefore selected a pilot agency - the government audit office of each country – on which to test our methodology. These audit offices have well-defined and comparable roles and responsibilities and all operate sites of relatively small size when compared with the whole government domain. For our research we selected the audit offices of Canada (CA), the UK and the USA as the three major English-speaking countries as well as New Zealand (NZ) and the Czech Republic (CZ), both to include smaller sites in the sample and to show that the methodology is language independent.

During October 2005, all documents and associated links were collected from the websites using the Nutch 1.6.0 crawler[2]. We started each crawl from the home page of the associated audit office and collected all pages up to a depth of 18. The crawl was restricted to the audit office domain and the crawler was configured to allow for complete site acquisition.[3] The composition of our crawl, according to page type, is shown in Table 2. In addition to crawling, we acquired the URLs pointing to these websites using Google's reverse lookup capability. It should be noted that the link data provided by Google is limited to the top 1000 results – affecting the highly linked pages in our dataset. Also Google inlink information has been previously shown to be of limited reliability[4].

---

[1] For example the following institutions at University College London (http://www.ucl.ac.uk) use different form of URL: http://www.cs.ucl.ac.uk/ (Department of Computer Science) http://www.ucl.ac.uk/spp/ (School of Public Policy)

[2] http://lucene.apache.org/nutch/

[3] The non default parameters being a 5s delay between requests to the same host, and 10,000 attempts to retrieve pages that failed with a soft error. Our crawler followed http links to files, skipping unparseable content.

[4] http://blog.searchenginewatch.com/blog/050128-134939

Table 2. Sizes of audit offices' websites.

Country	Google	Yahoo!	MSN	Crawl	doc	pdf	dynamic
CA	139,000	13,800	30,913	12,730	2	725	596
CZ	739	632	1,059	926	11	464	356
NZ	558	1,230	1,115	836	1	352	0
UK	32,200	6,980	10,325	4,027	78	1,411	144
US	433,000	64,400	72,010	19,625	0	4,897	4,140

Table 2 enumerates the size of the crawl for each website, together with the estimated size of these sites as provided by Google, Yahoo! and MSN. It can be seen that Google's estimates are usually larger than MSN's, which are usually larger than Yahoo's. The source of this discrepancy is unknown but is probably connected to the size of the underlying index as well as the associated estimation technique.[5] Reliable figures for the actual size of e-government websites are difficult to obtain. We are unaware of any government figures pertaining to this. While the number of pages we crawled is on the low side, personal correspondence with contacts at the UK audit office suggests that (at least for the UK) our crawl was exhaustive.

4. EXPERIMENTAL RESULTS

We analyzed the five datasets (CA, CZ, NZ, UK, US) for the audit offices of Canada, the Czech Republic, New Zealand, the UK, the USA, respectively. We examined both the internal structure and the external connectivity. Section 4.1 summarizes the internal structure, which is related to navigability, and Section 4.2 summarizes external connectivity, which we interpret as a measure of an institution's nodality.

4.1 Internal Structure

The internal structure of a website can have a significant effect on its navigability when users only use the hyperlinks to navigate the site. For the five datasets, we examined (i) the size of the connected components, (ii) the average distance between randomly selected pairs of nodes and their distribution, and (iii) the distribution of links within a site. These three properties can clearly affect navigability and are discussed in detail below.

4.1.1 Connected Components

Major advantage of hyperlinked environments is that they permit the user to navigate from one document to another and reach related documents. Which documents a user can reach on a website is primarily determined by the links that are included in each Web page. The drawback of not having any links from a page becomes obvious when we realize that a substantial proportion of users arrive via a search engine and therefore cannot use the back button to continue exploring the site. This seems to be supported by preliminary results of our user studies that show users starting navigation not from the top page but from deep inside a site (after arriving from a search engine) and then performing non-trivial navigation.

Since a user can enter a site at an arbitrary location provided by a search engine, a very simple assumption is that a good website should provide the possibility to navigate to any other page on the site via its hyperlinks, in other words that there is a path between any two pages. In fact, Broder et al [27] established that for the Web a path exists only for 25% of all pairs of nodes, i.e. in 75% of cases it is no possible to navigate between two random pages.

Broder et al also described the structure of the Web graph. At the centre is a *strongly connected component (SCC)* with a path between every pair of nodes. The *IN component* contains nodes that have a path to nodes in the SCC but not from the SCC. In the same way, nodes that are only reachable via a path from nodes in SCC but not conversely form the *OUT component*.

Table 3. Percentage of pages in strongly connected (SCC) and in OUT component, for both entire site as well as "navigable" site (i.e. without .pdf, .doc and image files) and percentage of documents removed by this filtering operation.

	Whole Site		Filtered site		%age of
Country	SCC	OUT	SCC	OUT	pdf+doc
CA	47	53	50	50	6
CZ	35	65	71	29	51
NZ	52	48	90	10	42
UK	34	66	54	46	37
US	52	48	70	30	25

In terms of a user navigating a site via its links, any node contained in the SCC is reachable from any other node in the SCC, although the path from one node to another may be long (see Section 4.1.2). In contrast, nodes in the OUT component "trap" a user since it is not possible to reach the nodes in the SCC from there. None of our datasets contain an IN component because the crawls started at the top page of the site that is part of the SCC. Table 3 summarizes the sizes of the SCC and OUT components in our datasets. Since our crawler does not parse .pdf and .doc files for hyperlinks, these documents are always in the OUT component. Thus, websites that provide many reports in pdf or doc format may appear to have a small SCC. We therefore repeated our analysis of the datasets without pdf, doc and image pages so that only navigable pages were included in the analysis.

We expect that sites with a large SCC will be easier to navigate than sites with a small strongly connected component because a user will not get trapped on a page that does not provide any link back to the central core of the site. Our analysis of the entire site reveals that approximately 50% of the US, NZ and CA sites are strongly connected compared with only 34% for the UK and CZ..

When only navigable content is considered, the SCC increases according to the number of pdf and Word documents on the site. New Zealand's ranking improves dramatically, with almost 90% of its site forming part of the SCC. The good performance of the US is especially noteworthy as it is the largest site of all.

There are several potential drawbacks to comparing these numbers. One open question is whether some documents that could not be parsed for links by our crawler (for example pdf and

[5] http://blog.searchenginewatch.com/blog/050128-134939

Table 4. Navigability related properties of audit offices' websites.

Country	Directed Diameter	Average Directed Distance	Median Directed Distance	Normalized			%age of Unreachable Pairs	
				Directed Diameter	Average Directed Distance	Median Distance	Whole Site	Navigable Content
CA	12	4.7	5	2.92	1.14	1.22	53%	50%
CZ	6	3.8	4	2.02	1.28	1.35	65%	29%
NZ	8	3.3	3	2.74	1.13	1.03	48%	10%
UK	22	5.4	5	6.10	1.50	1.39	66%	46%
US	23	5	5	5.36	1.16	1.16	48%	30%

doc files) should be excluded because these are by definition in the OUT component. Another question is how strongly the size of a website influences the size of the SCC. For smaller sites, it is relatively easy to ensure a SCC of 100%, which may explain why CZ and NZ have such large SCC's after filtering. In contrast, this is much bigger challenge for a site like the audit office of the US with almost 20,000 pages. It seems clear that some normalization is needed to account for the size of a website. However, we are uncertain what form this normalization should take.

4.1.2 Distance

While the strongly connected component indicates what percentage of the site can be accessed by navigating the link structure, it does not reveal the number of clicks needed to move from one node to another. It is therefore interesting to measure the distance, in number of hyperlinks followed, in order to navigate between two randomly selected nodes of a website.

For the following calculations we establish the longest shortest path between a random node and all other nodes. This is the longest path a user would have to follow to get from one node to another.

A worst case measure of distance is the directed diameter of the site, which is defined as the longest of all calculated shortest paths. Perhaps a more useful measure is the average distance, which is defined as the average over all the longest shortest paths of each node. We also quote the median as this measure is less susceptible to extreme outliers. Table 4 enumerates these values for each dataset. The path length does not change significantly when calculated for the whole site and only for the navigable content. This is because unreachable paths are ignored in these calculations. However, obviously the percentage of unreachable pairs can shrink dramatically. It is interesting to observe that the percentage of unreachable pairs is almost equal to the percentage of the OUT component. This indicates that there are almost no paths between pairs of nodes in the OUT component, i.e. the number of reachable pairs consists almost entirely of nodes in SCC. This implies that once a user enters the OUT component, there is not only no way back, but there is very little chance of navigating to another page.

The average directed distances of the websites in our sample are much lower than the ones that Albert et al [35] calculated for the Internet, indicating that indeed there is a higher degree of connectivity within a managed website than for the Internet as a whole. However, as with the SCC, it is clear that a small website will tend to have a smaller diameter and an average distance than a large website. Therefore in order to meaningfully compare the values, they should be normalized. We decided to apply a normalization factor equal to the logarithm of the size of the website. This is based on the models of [35] and [39] where the diameter grows linearly with logarithm of the size of the network.

Looking at Table 4, we observe that the normalized average distance for CA and US is now comparable with NZ. The normalized average distance for CZ has worsened, while the UK's normalized average distance remains the worst.

The average distance does not reveal the distribution of longest shortest paths. Figure 1 shows the cumulative distribution of path lengths for each data set. The asymptotic value of each curve has been normalized to reflect the different percentage of nodes that are unreachable for each dataset.

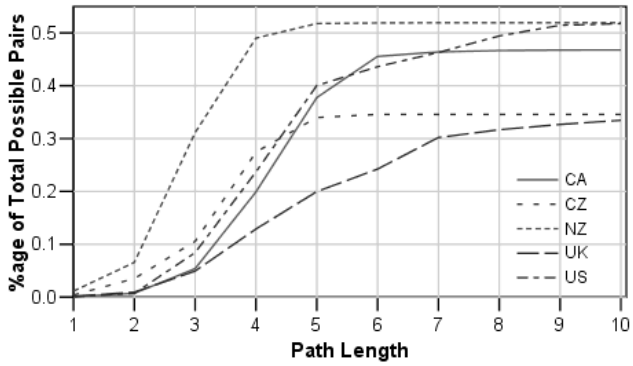

Figure 1. Cumulative sum of number of pairs of pages that have a path between them of less than a certain length. The *x*-axis represents the path length. The *y*-axis represents the fraction of all possible pairs of pages in the website that are connect by a path of length less than x.

Figure 1 shows that for those nodes that are reachable, the majority can be reached by six or less hyperlinks. While this value might seem rather small, Huberman et al [23-25] found that an average user follows only 4 links within a site. For a path length of four or less, we see that (i) for NZ, most of the reachable nodes are accessible, (ii) for CZ, over 50% of reachable nodes are accessible, (iii) for CA and US, somewhat over a third of reachable nodes are accessible, and (iv) for the UK, about a third of reachable nodes are accessible. This is particularly poor for the UK, since not only does it have the lowest accessibility for path lengths of four or less, but this is compounded by the fact that the percentage of reachable nodes is also the lowest.

4.1.3 Average Degree, Degree Distribution and Centralization

The previously discussed measures all rely on the existence of links between documents. One could argue that the more links exist, the more likely it is that there exists a path between two randomly selected pages. In order to measure how densely connected a network is we can measure the average degree of a node, that is the average number of links pointing to or from a node in a network. The higher the average degree, the more links exist in a network. As we use the average, this measure can be compared across networks of different size. In the same way, we can measure the indegree and outdegree distributions. However, as we analyze a website and its internal links only, every link going out from one page is received by another page within our graph. Therefore, the average indegree equals the average outdegree. In accordance with our earlier findings, the values in Table 5 underline that sites with a smaller normalized average distance tend to have more links.

Table 5. Average degree per page (sum of incoming and outgoing links), standard deviation from average and median of degree. Degree closeness centralization $C_d(G)$ both for the whole site crawled and for navigable content only.

	Degree		Degree Closeness Centralization	
Country	**Average (std. dev.)**	**Median**	**Whole Site**	**Navigable Content**
CA	20.0 (105.0)	3	0.341	0.341
CZ	7.7 (18.5)	2	0.427	0.507
NZ	19.9 (47.6)	14	0.607	0.789
UK	13.3 (57.3)	2	0.288	0.229
US	23.8 (277.9)	10	0.509	0.515

Many links do not necessarily guarantee a small average distance. Although the Web is a sparsely connected graph, it has been shown that it exhibits the characteristic of a small world [35][27][37], i.e. that most nodes can be reached from a random node by a comparatively small number of clicks. This is possible because links are unevenly distributed across nodes: there is a small number of nodes with a very high number of links while at the same time there is a large number of nodes with only a few links. The distribution of links follows a power law. For the Internet Albert et al [35] calculated a Zipf-exponent of 2.1 for the indegree distribution and of 2.72 for the outdegree distribution.

Figure 2 plots the distribution of indegree and outdegree for the websites in our sample. In the log-log plot, a straight line from the upper left corner to the lower right corner indicates a power law distribution. The distribution is stronger for larger sites, but even for the small sites, the indegree distribution roughly follows a power law. This is not the case for the outdegree distribution.

It is important to note that we only consider the distribution of links that are internal to the site, that is link source and link target are pages within the website of the audit office. In order to meaningful compare our distribution with distributions found for the Web as a whole, we would have to consider the external links as well.

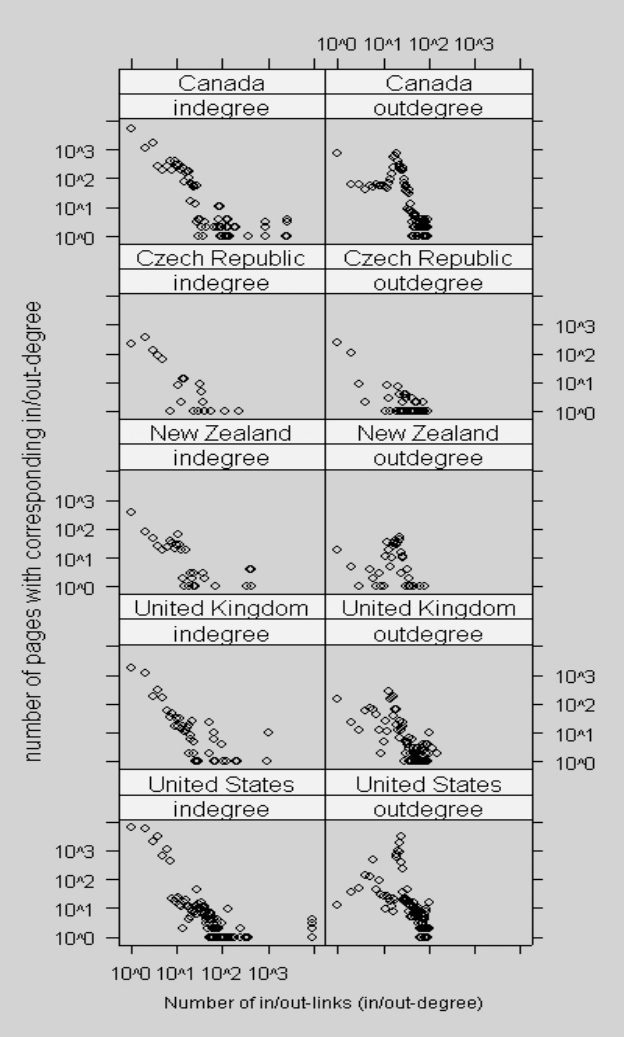

Figure 2. Distribution of in and outdegree for internal links in audit office websites as seen on a double logarithmic scale.

Although there is considerable variation, for inlinks, the five sites follow the general model of the Web, with a few nodes acting as internal authorities. While a few sites receive the majority of links, the websites differ in the amount to which they are centralized in their organization.

We use the *closeness centralization coefficient denoted* $C_d(G)$ [17] to compare centralization of sites. Let G be an undirected graph with n nodes representing our dataset where $V(G)$ and $E(G)$ are sets of vertexes/pages and edges/links respectively. Let $d(v,i)$ be the distance from node v to node i and S_n^* a star-network of n nodes. The closeness centrality of a node v *denoted* $c_d(v)$ is defined as:

$$c_d(v) = \frac{n-1}{\sum_{i \in V(G)} d(v,i)}$$

For the closeness centralization of a whole network we first calculate the overall variation in closeness centrality scores for all nodes, *denoted* $v_d(G)$:

Table 6. Inlink analysis of audit offices' websites.

	Number of External Inlinks		% of Pages Receiving Links	% of Links to Top Page	Number of Hosts		% of Government Hosts	Number of Different	
	Total	Normalized			Total	Normalized		Governments	Audit Offices
CA	647	32.4	0.9%	47%	212	10.6	25%	15	11
CZ	89	18.5	1%	90%	55	11.5	20%	10	8
NZ	39	12.2	0.1%	100%	34	10.6	44%	10	7
UK	2515	66.9	7.6%	57%	456	12.1	18%	25	20
US	8594	46.5	13.5%	9%	1905	10.3	12%	7	2

$$v_d(G) = \sum_{v \in V(G)} | c_d(v) - \max_{j \in V(G)} c_d(j) |$$

This variation is then set into relation to the variation in a network of the same size with maximum centralization, that is a star-network:

$$C_d(G) = \frac{c_d(G)}{c_d(S_n^*)}$$

From the definition the coefficient C_d takes values between 0 and 1 with one for a most centralized star–network.

The centralization measure can assist in the interpretation of our earlier measure of the datasets' diameter and average distance. This is because there are basically two ways to achieve a low distance between pages as well as a large strongly connected component – either a website can be built by simply linking all pages directly to a hub page, or by linking every page to different relevant pages on the site. While both methods will reduce the average distance between pages, the former results in a highly centralized network, while the latter creates a much more distributed network. We note that New Zealand is the most centralized network. However, further analysis is needed to determine the nature of this centralization.

4.2 External Connectivity

The previous sections analyzed the internal structure of the websites in our sample. It is hoped that analysis of the internal structure can provide a quantitative measure of the navigability of the websites. However, such an analysis does not relate to the "nodality" of political science, a concept that is akin to the hubness and authoritativeness of a site. To do so, requires broadening the analysis beyond the website itself, by establishing what are the links between the websites in our sample and other websites. Note that in the following analysis, inlinks and outlinks refer to links that come from or go to a node that is not part of the website under study.

4.2.1 Inlink Analysis

Following the common interpretation of a link as an endorsement, we look at number and type of links a website receives. The total number of inlinks, as reported in Table 6, can be interpreted as an indicator of a site's visibility or authority. We observe that the number of inlinks is very high for the UK and US.

However, the total number of inlinks does not reflect the fact that some countries have a very much smaller population than others. We therefore believe that it is appropriate to normalize the number of inlinks to a website by the estimated size of the internet population of the associated country. Consider the audit offices of large countries such as the United States, which have very high Internet penetration, and those audit offices that operate in smaller countries where less people have access to the Internet. Therefore we normalized the total number of incoming links to a website by the number, in millions, of Internet users in the country [34]. It is worth noting that this normalization is particular to governments online, in contrast to company websites for example. Although everyone can link to a national audit office's website, regardless of whether they are a citizen of that particular country, the very function of these institutions is primarily of interest to citizens of that country. This is supported by the finding that the majority of external inlinks originate from websites within the same country domain (see Figure 3). After normalization, we observe that the UK is ranked highest, i.e. is most authoritative, followed by the US and CA.

A site that offers a variety of useful information is more likely to receive links to many of its pages, not just to its home page. We therefore measure the proportion of pages on a website receiving external links. We argue that the higher this value is, the more useful information is likely to be offered by the site. Additionally, we determined the proportion of external links pointing to the home page. We believe that a low proportion of links to a home page is better, as this probably indicates that external sites are pointing to specific, useful information on the site rather than pointing to an audit office as an institution. By this measure, we observe that all pointers to NZ are generic, while only about 9% of inlinks to the US point to the home page.

The total number of inlinks does not reveal their distribution. For example, all inlinks might originate from a single external site. We believe it is better if the inlinks come from a variety of external sources. We report the number of different hosts pointing to each site. Again, we normalize this measure by millions of Internet users in country. Interestingly, although the US exhibits by far the largest number of hosts the normalized measure suggests that both the UK and even CZ are in fact linked to by more different websites than the North American audit offices.

As we are interested in e-government, the links between different national governments are of interest as well as the external links within the national government of the country. These are tabulated in Table 6 and graphically illustrated in Figure 3. We see in Figure 3 that the websites in the sample differ in the proportion of inlinks originating from other government domains. Still, it can clearly be seen that the majority of websites pointing to an audit office is non-governmental, indicating that the relevance of these institution extends beyond the institutionalized political system. In fact, as one would hope, business interest is strong – about a third of inlink sources are commercial. Furthermore, national audit offices really seem to be 'national'

Table 7. Outlink analysis of the audit offices' websites.

	Number of External Outlinks		Number of External Pages Linked to	% of Pages Containing External Link	Number of Hosts		% of Government Hosts	Number of Different	
	Total	Normalized			Total	Normalized		Governments	Audit Offices
CA	5516	0.43	267	39%	98	0.0077	67%	4	1
CZ	21	0.02	8	1%	8	0.0086	50%	4	3
NZ	526	0.63	96	50%	79	0.0945	52%	16	13
UK	886	0.22	568	4%	308	0.0765	44%	48	43
US	243	0.01	190	<1%	123	0.0064	35%	1	1

institutions – at least half of all websites pointing to an audit office are located in the same country. This adds relevance to the normalization by national Internet population that we applied earlier.

4.2.2 Outlink Analysis

The number of outlinks from a site is related to its 'hubness' [30]. In political science terms, hubness could be seen as a measure of the extent to which a organization 'collects' information from the outside world, by providing users with links to other sources of information.

Governments tend to perceive themselves as the ultimate authority on information. Therefore we assume it is less likely for government websites to function as hubs and to point to other sources of information, except perhaps to other governments.

Table 7 provides a comparison of outlink statistics for our five datasets. There are substantial differences in the number of external outlinks across sites. While it was felt appropriate to normalize the external inlinks of each site by the respective sizes of their internet populations, such normalization does not seem appropriate for outlinks as the creation of an outlink is not performed by individual users. Instead, we propose to normalize by the size of each site, to provide the number of outlinks per page. Surprisingly, after normalization, we see that NZ exhibits the highest number of outlinks per page (63%). This may be misleading, since it is quite common to link to generic sites such as "Adobe Acrobat Reader". This is confirmed when we consider column 4 of Table 7, the number of unique external pages pointed to. Here, we see that while the total number of outlinks for NZ is 526, the total number of different outlink targets is only 96.

A similar change can be observed for Canada if one considers the number of hosts pointed to instead the total number of outlinks. Although Canada links heavily outside, it does not point to many different websites. In fact, a closer examination of Canada's outlinks revealed that the majority of these links have one single target – the website of the Canadian government. In contrast, the National Audit Office of the United Kingdom is heavily interlinked with foreign governments, unlike Canada and the United States. While a considerable share of the outlinks point to government websites, audit office link also to other sources.

We can distinguish the websites linked to by the audit office's according to their top-level domain. The North American countries are primarily inward looking with few links to websites in foreign countries. Here clearly the UK takes the lead. Another interesting result is that most audit offices are obviously not afraid of linking to commercial sites.

5. DISCUSSION

We have performed a preliminary comparative study of government audit offices on Canada, the Czech Republic, New Zealand, the UK and the USA. This study was based on an examination of the structural characteristics of these websites. Website properties based on usage statistics were not considered due to the difficulty of acquiring such data. Despite this lack of usage information, it appears that structural information is informative as to the quality of the websites.

The structural characteristics examined form two categories, internal structure that is indicative of the navigability of a site, and external structure that is indicative of the nodality (hubness and authority) of a site. A further subdivision may be made with respect to nodality. We use the distinction between hubs and authorities to distinguish between nodality in terms of collecting information (hubness) and nodality in terms of disseminating information (authoritativeness).

A variety of properties were examined – diameter of the site, average distance between nodes, largest strongly connected component, and the number of external inlinks and outlinks. These properties have been well studied within computer science, the Web and the social sciences. Nevertheless, our comparative analysis revealed some shortcomings of these properties, due to the diversity of the websites and countries under investigation. It is clear that in many cases, these properties need to be normalized to account for the size of a website and/or the size of the national Web. In the latter case, this was approximated by normalizing for the internet penetration in each country. Other normalizations are possible and future work will consider such issues as the variation in gross domestic product of each nation. Furthermore, a simple count of the number of external inlinks and outlinks can be misleading. At the very least, it is necessary to adjust the results to only consider unique source or target pages/hosts. The US and Canada emerge as the most internally connected and navigable sites in relation to their size, with the UK scoring the lowest on this dimension. The smallest sites, the Czech Republic and New Zealand, score highest in absolute terms, having the largest strongly connected component, the lowest average and median directed diameter, the lowest percentage of unreachable pairs and the highest degree of centralization (taking only navigable content into account, as we believe should be the case). For the larger sites, the US and Canada score generally better for navigability than the UK across the same measures, in spite of being far larger than the UK site. When we take size into account (by normalizing for the number of pages), the UK's position as a laggard becomes most marked, with the highest directed diameter, directed distance and percentage of unreachable pairs of all the sites, including the Czech Republic and New Zealand. To summarize the results with

respect to navigability, internal structural characteristics reveal that the US and CA rank highly with regard to these navigability measures, while the UK is seen to be lagging.

In contrast, with respect to nodality, the UK and the US emerged as the clear leaders. In terms of incoming nodality (or authority), the UK scored most highly with respect to the normalized number of external inlinks, and was beaten only by the US with respect to the proportion of pages receiving links. These two were considerably more authoritative in this sense than Canada, which in turn eclipsed the two smaller sites. The US and the UK received links from a greater number of hosts than the other sites, with a smaller proportion of governmental hosts. The UK had far more links from different governments (probably reflecting the NAO's hosting of a site for an international association of audit offices); the US was lowest here with only 7, perhaps highlighting the somewhat insular nature of US public administration. When it comes to outgoing nodality (hubness), which is a measure of the 'willingness' of a site to link to information in other domains, we found considerable variation, with Canada having by far the greatest number of outlinks. New Zealand and then the UK have a significantly higher number of external outlinks when normalized by the size of the website. Although Canada linked heavily outside, the majority of links are within the Canadian government and it was the UK that emerged as linking to the highest number of hosts (over twice as many as its nearest rival, the US), the highest number of different governments (four times as many as its nearest rival, New Zealand) and by far the highest number of audit offices. The US and the UK appear to be the most 'authoritative' sites, i.e. the most effective disseminators of information. The UK and New Zealand seemed to be the most effective collectors of information (hubness) from the outside world. Overall, therefore, we might deem the UK the most 'nodal' of our audit offices online.

Our comparative study affirms previous qualitative studies, for example in their findings of the superiority of Northern American sites. However, we have done so by using quantitative measures instead of qualitative assessment. Our study has also yielded novel results. By distinguishing between internal and external connectivity, we have broadened our analysis beyond the website itself and have established a first quantification of nodality and therefore how an institution's website relates to its surroundings.

We believe these metrics offer the possibility to provide a more sophisticated and meaningful evaluation of the web structures of government than any of the existing studies outlined in the first section. When applied at the 'whole government' or policy sector level, they offer the possibility to rigorously assess the accessibility of government information along two key dimensions: navigability and nodality. This study represents, as far as we know, the first attempts to quantitatively measure either of these dimensions and the first attempt to apply these techniques systematically to the web structure of government. Because these measures are non-obtrusive, data can be collected relatively cheaply (more so, at least, than any user metrics) without transgressing ethical boundaries or seeking multiple permissions although a web crawler has to be configured in a way that puts no irresponsible pressure onto the bandwidth and availability of information providers [40]. Our aim now is to move beyond the relatively modest research subject of audit offices to larger departments (such as finance ministries or foreign offices) and then to governmental domains.

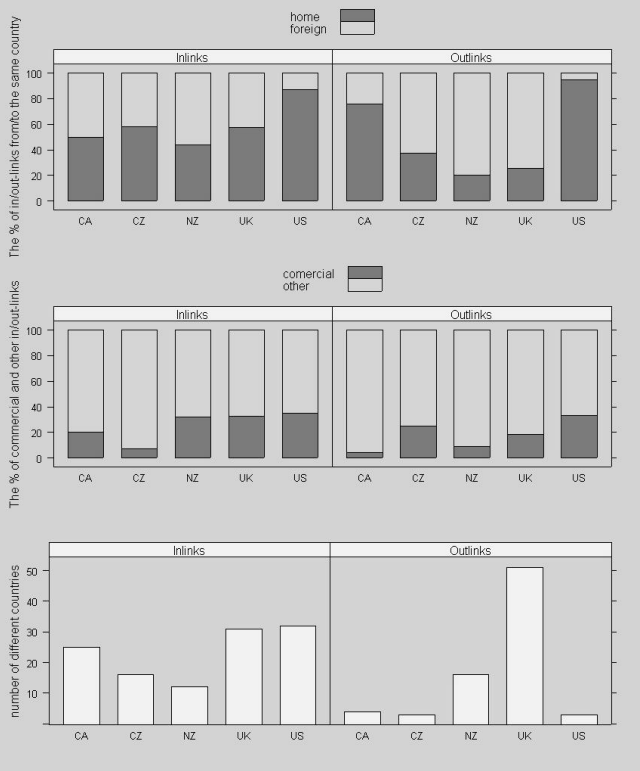

Figure 3. A comparison of the audit offices according to (i) percentage of hosts from the home country, (ii) percentage of commercial hosts and (iii) number of different countries appearing in the list of hosts.

The next stage for future research will be to compare our structural measures against results of our lab based user study that we just finished and also against user metrics, collected via mystery shopping exercises, opinion surveys and usage statistics, in order to verify that sites or communities which emerge as 'healthy' in terms of navigability and nodality also score well when experienced by users.

6. ACKNOWLEDGMENTS

Thanks to the National Audit Office of the UK and the Web manager of the School of Public Policy, Aaron Crompton, for providing us with data to verify our crawls. We also want to express our gratitude to Vladimir Batagelj and Andrej Mrvar for developing Pajek, their excellent program for large scale network analysis. We would like to thank the Cambridge-MIT Institute for financial support. This work was supported in part by the National programme of research (Information society project 1ET100300419).

7. REFERENCES

[1] Accenture *'Leadership in Customer Service: New Expectations, New Experiences'*, The Government Executive Series, April 2005

[2] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, M. Gori *'Focused Crawling using Context Graphs'*, in 26th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, 10–14 September 2000, pp. 527–534.

[3] A. Caldas *'On the Web Structure and Digital Knowledge Bases'*, 2004.

[4] Cap Gemini, '*Online Availability of Public Services: How is Europe Progressing? Web-based survey on electronic public services, Report of the fifth measurement'*, European Commission Directorate General for Information Society and Media, 2004.

[5] D. Dalziel, '*Government online: A multi-country study of e-government usage*', World Association of Research Professionals, 2004.

[6] C. C. Demchak, C. Friis, T.M. La Porte, '*Webbing governance: National differences in constructing the public face'*, in G. D. Garson (ed.) Handbook of Public Information Systems (New York: Marcel Dekker Publishers), 2000.

[7] N. Eiron,, K. S. Mccurley '*Locality, Hierarchy, and Bidirectionality in the Web'*, Workshop on Web Algorithms and Models, 2003.

[8] W. H. Dutton, C. di Genarro, A. Millwood, '*The Internet in Britain: The Oxford Internet Survey (OxIS)*, Hargrave, May 2005,

[9] I. Graafland-Essers, E. Ettedgui, '*Benchmarking E-Government in Europe and the U.S*', Rand Research, 2003.

[10] M. Hindman, K. Tsioutsiouliklis, J. A. Johnson, '*Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web*', 2003.

[11] C. Hood, '*The Tools of Government'*, Macmillan, 1983.

[12] C. Hood, H. Margetts, '*The Tools of Government in the Digital Age'* (London: Palgrave), 2006.

[13] T. M. La Porte, C. C. Demchak,, C. Friis, '*Webbing Governance: Global Trends across National Level Public Agencies'*, Communications of the ACM, January 2001.

[14] Taylor Nelson Sofres, '*Government Online: An international perspective*', Annual Global Report, 2003.

[15] United Nations, '*World Public Sector Report 2003: E-Government at the Crossroad*',*s* Department of Economics and Social Affairs (New York: United Nations), 2003.

[16] D. West, '*Digital Government: Technology and Public Sector Performance'*, (Princeton University Press), 2005.

[17] W. Nooy, A. Mrvar, V. Batagelj, '*Exploratory Social Network Analysis with Pajek*', Cambridge University Press, 2005.

[18] H. W. Park, '*Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web*', Connections 25(1):49-61, 2003.

[19] M. Thelwall '*Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites'*, JASIST, 53(12):995–1005, 2003.

[20] L. Page, S. Brin, R. Motwani, T. Winograd '*The PageRank Citation Ranking: Bringing Order to the Web'*, Tech. Rep., Stanford Digital Library Technologies Project, 1998.

[21] K. Bharat, B. W. Chang, M. R. Henzinger, M. Ruhl '*Who Links to Whom: Mining Linkage between Web Sites'*, in ICDM, pp. 51–58, 2001.

[22] X. He, H. Zha, C. Ding, H. Simon '*Web document clustering using hyperlink structures'*, CS&DA 41:19-45, 2001.

[23] B. A. Huberman,, P. Pirolli, J. E. Pitkow, R. M. Lukose, '*Strong Regularities in World Wide Web Surfing*'. Science, 280, pp 95-97, 1998.

[24] L. Adamic, B. Huberman 'The Web's hidden order', Communications of the ACM, vol. 44, no. 9, 2001.

[25] B. Huberman, L. Adamic, '*Growth dynamics of the World-Wide Web'*, Nature, vol. 399 pp. 130, 1999.

[26] S. Brin, L. Page, '*The anatomy of a large-scale hypertextual Web search engine*'. Computer Networks and ISDN Systems, vol. 30, no. 1–7, pp 107–117, 1998.

[27] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, '*A. Graph structure in the web: Experiments and models*', 9th WWW, 2000.

[28] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, '*Automatic resource list compilation by analyzing hyperlink structure and associated text'*, in Proceedings of the 7th International World Wide Web Conference, 1998.

[29] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, J. Kleinberg, '*Mining the Web's Link Structure'*, Computer, vol. 32, no. 8, pp. 60–67, 1999.

[30] D. Gibson, J. M. Kleinberg, P. Raghavan, '*Inferring Web Communities from Link Topology'*, in UK Conference on Hypertext, pp. 225–234, 1998.

[31] J. M. Kleinberg, '*Authoritative sources in a hyperlinked environment'*, Journal of the ACM, vol. 46, no.5, 604–632, 1999.

[32] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins '*Trawling the Web for emerging cyber-communities'*, Computer Networks (Amsterdam, Netherlands), vol. 31, no. 11–16, pp. 1481–1493, 1999.

[33] G. W. Flake, S. Lawrence, C. L. Giles, F. Coetzee, '*Self-Organization of the Web and Identification of Communities'*, IEEE Computer, vol. 35 no. 3 pp. 66–71, 2002.

[34] International Telecommunication Union '*Internet indicators: Hosts, Users and Number of PCs*', 2004.

[35] R. Albert, H. Jeong, A. Barabási, '*Diameter of the World-Wide Web*'. Nature, 401 (September 1999), 130.

[36] S. Zhou, R. J. Mondragon, '*Accurately modeling the Internet topology'*, Physical Review E, vol. 70, no. 066108, the American Physical Society, 2004.

[37] J. Laherrre, D. Sornette., '*Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales'*, The European Physical Journal B - Condensed Matter, 2(4), pp 525-539, 1998.

[38] P. Ingversen, '*The calculation of Web impact factors*', Journal of Documentation, 4(2), pp 236-243, 1998.

[39] L. Lu. '*The Diameter of Random Massive Graphs'*, in Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, 2000.

[40] M. Thelwall, D. Stuart, '*Web crawling ethics revisited: Cost, privacy and denial of service'*, to appear in Journal of the American Society for Information Science and Technology.