Estimating the Size of Political Web Graphs

Robert Ackland*
Centre for Social Research
Research School of Social Sciences
The Australian National University
Canberra, ACT 0200, AUSTRALIA
Final version: 25th April, 2005

The web pages and hyperlinks that form the World Wide Web can be viewed as nodes and edges in a directed graph (the *web graph*). In this paper, snowball sampling is used to estimate the size of web graphs inhabited by different political parties using data for 117 political parties from Austria, France, Germany, Italy, UK and Australia. It is found that mainstream left parties inhabit the largest political web graph, while the web graph for parties from the mainstream right is the smallest. Italian parties are found to be the most outward looking in terms of online networking with each party inhabiting a web graph that is 10 times larger than the average. In contrast, the web graph inhabited by the average Australian political party is less than half the size of the average for all countries.

Keywords: snowball sampling, political parties, networking, link analysis, WWW

*Postal address: Research School of Social Sciences, Coombs Building #9, Canberra ACT 0200, AUSTRALIA. E-mail: robert.ackland@anu.edu.au. URL: http://acsr.anu.edu.au/staff/ackland. Phone: +612 6125 0312.
I would like to thank Joel McDonald for programming assistance and Natalie Cooper for research assistance. This research has been supported by the Australian Research Council.

1 INTRODUCTION

Over the past 10 years, the Internet has become an important tool for communication, with individuals and organizations increasingly using e-mail and the World Wide Web (WWW) for day-to-day personal and professional interactions. In addition to its role as a new information and communication medium, the WWW is also highly effective in facilitating the formation and maintenance of networks and this has particularly been evident in areas of politics and social movements. Researchers are beginning to use web data to study this behavior and are using methods and technologies that information scientists have developed for the analysis of web data. For example, Hindman et al. (2003) use information retrieval methods to identify and characterize political communities on the WWW and show that based on the link structure within these communities, a small number of web sites dominate the provision of political information. Thelwall (2004a) provides an in-depth coverage of the use of purpose-built web crawlers for research and includes several social science applications. The question to be addressed in the present paper is: are the methods developed by information scientists such as information retrieval and web mining appropriate for social scientists who want to study networking behavior on the WWW?

The contention motivating the present paper is that the methods used by information scientists need to be adapted for use by empirical social scientists. In particular, while the potentially huge scale of web data sets is not a concern for information scientists (and in fact the existence of vast data sets is often the main motivation behind development of methods and technologies in the information sciences), large volumes of data can pose a serious problem for social science research. Automatic methods for classifying and characterizing large web data sets will only go so far. For research into the networking behavior of different organizations on the WWW, for example, it is not enough to simply know that there is a hyperlink connecting

two organizations on the WWW (something that web mining tools can easily show). We need to know the contextual information surrounding the link, as that will give further indications as to what the link actually means. At a bare minimum, we want to know exactly what organization is behind each node on the web graph, and this will be impossible in a data set containing millions of web pages (the data set used by Hindman et al. 2003 contains almost 3 million pages).

In this paper it is argued that for web data to become useful for social science research, we need methods for constructing representative samples of web data. Data from the WWW are essentially network data. The web pages and hyperlinks that form the WWW can be viewed as nodes and edges in a directed graph and this characterization of the WWW as a *web graph* underpins quantitative web research by physicists, computer scientists, information retrieval specialists, and, increasingly, social scientists.

In this paper, it is proposed that adaptive sampling methods, whereby the sample design (the process by which the sample is selected) is dependent on information collected during survey, may be appropriate for constructing samples of web data for social science research. A particular type of adaptive sampling, snowball sampling (see, for example, Frank and Snijders 1994), is introduced and applied to the estimation of the size of political web graphs (defined as the parts of the web graph inhabited by different political parties). Snowball sampling is used to assess whether there are significant differences in the size of the web graphs inhabited by 117 political parties from following different party "systems" (far left, left, centre, right, far right, ecologist, regionalist) and countries (Austria, France, Germany, Italy, UK and Australia).

It is argued that the size of the different political web graphs gives information on the online networking behavior of the parties in different political systems: parties in relatively larger web graphs are more active in using the WWW to "reach out" and network with other organizations, compared with parties inhabiting smaller web graphs. It is found that mainstream left parties inhabit the largest political web graph (connecting to 396 other websites, on average), while the

web graph for ecologist parties is the smallest (containing an average of only 37 websites per party). At a country level, Italian parties are found to be the most outward looking in terms of online networking with each party inhabiting a web graph that is around 10 times larger than the average for all 117 parties. In contrast, the web graph inhabited by the average Australian political party is less than half the size of the average for all countries.

To the author's knowledge, this is the first time adaptive sampling has been used for studying networking behavior on the WWW. Previous social science research using web data has tended to either completely eschew the topic of sampling (Hindman et al. 2003, p.7 state that their method of data collection and analysis "does not require the use of sampling...it can catalog *all* Web pages easily reachable using certain online tools given specified constraints...") or has equated sampling to the use of a search engine to collect web data (Brunn and Dodge, 2001).[1] With the sampling approach proposed here, web data can be used in an analytical framework that is familiar to quantitative social scientists and allows for the identification of statistically significant online behavior amongst individuals or organizations.

In Section 2 political web graphs are defined, and an approach for estimating their size using snowball sampling is outlined. In Section 3, estimates of the size of political web graphs are presented using the data from 117 political parties. Section 4 presents conclusions, caveats and issues to be addressed in future work.

2 POLITICAL WEB GRAPHS

We can represent the hyperlinks between a set of web pages as a directed graph where pages are represented as graph nodes or vertices and hyperlinks between pages are represented as graph edges or arcs. Let $V = \{1, \ldots, v\}$ denote the set of v vertices (the *vertex set*) in the graph.

[1] Thelwall (2004a)[Chapter 2] discuss the process of taking random samples of web pages returned from search engines and the factors that may lead to the introduction of sampling bias.

Let W denote the edge set, a subset of V^2, and for convenience let W also contain all loops $\{(i,i) : i \in V\}$. The edge set is represented by the indicator variables $y = (y_{ij} : (i,j) \in V^2)$, where y_{ij} is 1 or 0 depending on the existence of a directional link from i to j. Define the *adjacency matrix* of the graph as Y, where ijth element of Y is y_{ij} (note: $y_{ii} = 1, i \in V$, by construction).

Definition 1. WEB GRAPH*: A web graph is a directed graph defined by a set of web pages V and an adjacency matrix Y indicating the hyperlinks between these pages.*

The WWW can be modeled as huge web graph (containing billions of nodes and edges) and researchers have used graph theory to estimate properties of the web such as its size and the distribution of web traffic and inbound hyperlinks (see, for example, Albert et al. 1999 and Barabasi and Albert 1999). However, these "macro" methods have generally been developed in the context of research in the areas of applied physics and computer and information sciences, and they are not necessarily appropriate for use in social science research using web data. In particular, social science research does not necessarily "scale" well to data sets containing billions of observations since social scientists will want to know the context of a link between two pages and at the bare minimum, will want to know something about the organizations or individuals who have placed the pages on the WWW. A web graph containing billions of observations, even if it could be constructed and analyzed, would not be of much use for studying the online behavior of far-left political parties, for example. Another problem with applying these macro methods to social science research is that they do not provide for statistical inference - an essential tool of empirical social scientists. In the context of political science research, for example, we want to be able to make statements such as "Political parties from country x or party system y exhibit online networking behavior that is statistically different from the average political party".

What is required is a method for constructing a "purposeful" sample of web data that contains web data for an entity of interest (e.g. political party) and other entities connected to

this party in cyberspace.

Definition 2. INITIAL SAMPLE*: The initial sample, S_0, is a set of web pages that represent entities of interest on the WWW.*

In the present paper, initial samples are constructed at the level of the party system (or ideological family) and the country. For party system i, S_0^i contains the homepages of all parties classified as being part of that system, while for country i, S_0^i contains the homepages of all parties from that country. Note that for reasons to be discussed below, S_0^i must contain more than one element.

Definition 3. PURPOSEFUL WEB GRAPH*: A purposeful web graph constructed for initial sample S_0^i is a directed graph defined by a set of web pages V^i that are connected to (either directly or indirectly) the web pages in S_0^i and the adjacency matrix Y^i indicating the hyperlinks between these pages.*

A purposeful web graph thus represents "slices" or "portions" of the WWW that are relevant to a particular topic of social science research. It is important to note that without access to the entire web graph (containing billions of pages) a purposeful web graph is not directly observable. While it is possible to use a web crawler to identify the web pages that the pages in S_0^i link to (i.e. the outbound links from S_0^i), it is impossible to know all the pages link *to* the initial sample (i.e. the inbound links to S_0^i). However, as shown below, we are able to use snowball sampling methods to estimate an important property of a given purposeful web graph V^i, it size, which is denoted v^i.

It should also be noted that the term "connected to" used in the definition of a purposeful web graph is open to interpretation and allows for the judgement of the researcher constructing the graph. In particular, we will generally want to ignore hyperlinks to pages that that are seen as being irrelevant to the social science phenomenon being studied (this is referred to as *topic*

drift in the information retrieval literature). For example a hyperlink from the homepage of a political party to `www.adobe.com` might be ignored because it is most likely that this link is designed to facilitate the downloading of the acrobat reader (for reading pdf files) and has no political meaning. A hyperlink between two pages is a necessary but not sufficient condition for the two pages to be part of the same purposeful web graph.

Definition 4. POLITICAL WEB GRAPH*: A political web graph is a purposeful web graph constructed for initial sample S_0^i that contains web pages of a political nature, e.g. the homepages of political parties.*

2.1 Snowball Sampling

With *conventional sampling* approaches, the process for selecting the sample does not depend on observed patterns in the data. Thus, the sampling frame can be established in advance of conducting the survey. With *adaptive sampling*, however, the sample design (the process by which the sample is selected) is dependent on information collected during the survey.[2] That is, the sample design can change in response to observed patterns in the data.

Thompson and Collins (2002) distinguish two broad types of adaptive sampling: link-tracing designs (that use social network information) and other designs that primarily use geographic information. The focus of the present paper is on link-tracing designs, also known as graph sampling designs, which is where investigators use links between observations (e.g. people) to find other observations to include in the survey. Link-tracing sampling designs are adaptive because the social connections between observations only become evident in the course of the sampling process. Snowball sampling is a particular type of link-tracing design that has been used to estimate the characteristics of hidden and elusive populations. Frank and Snijders (1994) use snowball sampling to estimate the size of the population of heroin users in Groningen. In

[2]For reviews of adaptive sampling methods see Thompson and Seber (1996) and Thompson and Collins (2002).

this paper, these methods are used to estimate the size v^i of a purposeful web graph V^i.[3]

Denote by A_j and B_j the subset of nodes linked to by (or "after") node j and linking to (or "before") node j, respectively. That is,

$$A_j = i \in V : y_{ji} = 1,$$

$$B_j = i \in V : y_{ij} = 1.$$

Row j of Y thus indicates A_j and column j of Y indicates B_j. The number of nodes (or sizes) of A_j and B_j are the out-degree and in-degree of node j and are denoted a_j and b_j, respectively. They are calculated as the row and column sums of Y:

$$a_j = |A_j| = \sum_{i=1}^{v} y_{ji},$$
$$b_j = |B_j| = \sum_{i=1}^{v} y_{ij}.$$

For any subset S of V denote as $A(S)$ and $B(S)$ the subsets of vertices after and before any of the vertices in S, respectively:

$$A(S) = \bigcup_{j \in S} A_j,$$
$$B(S) = \bigcup_{j \in S} B_j.$$

For a given initial sample S_0, the first wave of the snowball sample is given by: $S_1 = A(S_0) \cap \bar{S}_0$, where $\bar{S}_0$ is the complement of S_0 (all vertices in V not in S_0). The second wave is given by $S_2 = A(S_1) \cap \bar{S}_0 \cap \bar{S}_1$, and in general $S_i = A(S_{i-1}) \cap \bar{S}_0 \cap \ldots \cap \bar{S}_{i-1}$. The snowball initiated by S_0 is given by $S_0 \cup S_1 \cup \ldots \cup S_K$, where K is the number of waves of the snowball.

[3]Much of the following presentation is based on that in Frank and Snijders (1994). Note that in the following presentation, the superscript i denoting a purposeful web graph defined for a particular group of entities is omitted for clarity.

2.2 Estimation Of v

As mentioned above, it is not possible to construct a political web graph for a given initial sample of political homepages. However, it is possible to estimate the size of the web graph, v. Frank and Snijders (1994) outline two broad methods for estimating v. With *model-based estimation*, the edge indicators y_{ij} are modeled as realizations of an underlying stochastic process. In contrast, with *design-based estimations* the underlying unobservable directed graph is considered to be fixed i.e. the edge indicators y_{ij} are not random but are unknown and probability only enters the estimation via the sampling procedure. As this paper represents a very preliminary foray into using adaptive sampling methods for estimating the size of political web communities, attention is restricted to only one of the several model-based estimators discussed in Frank and Snijders (1994).

Following Frank and Snijders (1994), assume that the initial sample S_0 is a Bernoulli subset of V with selection probabilities α. The initial sample size $n = |S_0|$ is then binomial (v, α). The edge set W is further assumed to be a Bernoulli subset of V^2 with selection probability 1 for the loops and selection probability β for other edges. Thus, the diagonal of the adjacency matrix Y contains 1s, while the off-diagonal elements are *iid* Bernoulli(β) variables. The unknown parameters of the statistical model are v, α and β, but as with Frank and Snijders (1994), the focus of this paper is v.

Frank and Snijders (1994) show that a moment estimator of v, conditional on n, is provided by:

$$\hat{v}_1 = n + s(n-1)/r,$$

where r is the number of nonloop edges within S_0 and can be found as $r = |W \cap S_0^2| - n$ (since the number of loops is equal to the size of S_0, n) and s is the number edges from S_0 to the first wave of the snowball, that is, $s = |W \cap (S_0 \times S_1)|$. It is apparent that we need $r > 0$

- this is why the initial sample S_0 must contain more than one element. Note that for small initial samples, it may be that $\hat{v}_1$ cannot be estimated because $r = 0$.

The estimated variance of $\hat{v}_1$ is:

$$var(\hat{v}_1) = (n^2 - n - r)(n - 1)s(s + r)/nr^3.$$

3 ESTIMATES OF SIZES OF DIFFERENT POLITICAL WEB GRAPHS

The above methods are now used to establish whether there are significant differences in the size of political web graphs constructed for political parties from different party systems and countries. This will give insights into whether there are differences in the way parties from different political systems and countries are using the WWW as a tool for networking.

3.1 Construction Of Initial Samples (S_0)

The web data were collected using the `uberlink` software package (Ackland, 2005) that incorporates a GUI, web crawler, relational database and visualization tools and can be used to collect and analyze data pertaining to online networks of individuals and organizations. As discussed in Ackland and Gibson (2005), the web homepages for 117 political parties from six established Western democracies (Austria, France, Germany, Italy, UK and Australia) were identified. Using a range of sources from the literature on party classifications and expert advice, each party was classified into one of seven ideological or party families: far left, left, centre, right, far right, ecologist, regionalist. The distribution of the 117 parties across the countries and party families is shown in Table 1.

The subsets of political parties, classified according to party system and country of origin, comprise the initial samples (S_0) that are to be used to estimated the size of the different political

Table 1. Political parties in sample by country and party system

	Austria	France	Germany	Italy	UK	Australia	All
Far Left	2	6	1	2	2	3	16
Left	3	7	2	4	1	8	25
Centre	0	2	3	4	2	2	13
Right	4	4	5	5	1	9	28
Far Right	2	2	3	3	7	3	20
Ecologist	1	4	1	1	1	2	10
Regionalist	0	0	0	2	3	0	5
All	12	25	15	21	17	2 7	117

Note: Source - Ackland and Gibson (2005).

web graphs. It should be noted that the method of selecting the initial samples is such that they are almost certainly *not* Bernoulli samples. However, it is not clear that the degree of departure from Bernoulli sampling will be systematically different across different party systems and countries, and hence it is argued that the resulting estimates of sizes of political web graphs will be valid for comparisons across these dimensions (the objective of this paper). All the same, future research will focus on making adjustments for varying selection probabilities in the initial samples.[4]

3.2 Estimates Of Political Web Graph Size

In Table 2, initial estimates of political web graph size ($\hat{v}_1$) are presented.[5] The unit of measurement is *web page group* (or "site"), rather than web page. While the raw data collected by `uberlink` are web pages, the software provides the facility for pages to be aggregated into page groups, which are defined as aggregations of web pages that represent an entity of interest.[6] Thus all the pages of an entity of interest such as a political party or other organization can be represented as a single node in the web graph, rather than a node for each page. For the

[4]Frank and Snijders (1994)[p.66] refer to on-going research on this issue.

[5]NOTE TO REVIEWERS: Data and code for replicating the results presented in this paper are available on request from the author; on acceptance for publication, these resources will be made publicly available via URL link in the paper or another method appropriate for the journal.

[6]See Thelwall (2002, 2004b), Thelwall and Harries (2003) and Thelwall and Wilkinson (2003) for more on aggregating pages into groups or clusters using alternative document models (ADMs) based upon directories, domains and multi-domain sites.

present analysis, pages that share the same domain name, e.g. `www.alp.org.au`, are aggregated into a single page group. The problem with this approach is that personal homepages being commercially hosted by the same internet service provider will be aggregated into the same page group, even if they represent different entities of interest. In future work, page groups will be created in a more accurate manner.

When all parties are included in the initial sample, the estimated size of the web graph is 5726 page groups. Looking at each party system separately, mainstream left parties inhabit the largest web graph, estimated to consist of 9889 page groups; this compares with the much smaller web graph for regionalist parties of only 321 page groups. Note that the precision of the estimates of the web graph size for some of the party systems is quite poor - it is apparent that when $r = 1$, the standard deviation of $\hat{v}_1$ is approximately equal to $\hat{v}_1$.

The problem with the estimates of size of web graphs presented in Table 2 is that they are dependent on the size of the initial sample S_0, which varies greatly across the different party systems. Thus, the estimates $\hat{v}_1$ are not good indicators of the networking behavior of the individual parties. In the last column of Table 2, estimated web graph size per party in the initial sample ($\hat{v}_1/n$) is presented - this is the preferred indicator of networking behavior. Based on this measure, mainstream left parties still inhabit the largest web graphs (396 page groups per initially sampled party), while ecologist parties form a web graph that contains only 37 page groups per initially sampled party.

At a country level, Italian parties are found to inhabit the largest web graphs (491 page groups per initially sampled party) while Australian parties appear to be quite unmotivated with regards to online networking, inhabiting a web graph containing only 20 page groups per initially sampled party (compared with an average for 117 parties of 49).[7] Italian parties appear to have a much more "outward" focus with regards to their linking behavior and this is leading to a larger estimate of the size of the web graph they inhabit. Of the outbound links from the

[7]It was not possible to construct estimates of $\hat{v}_1$ for Austrian parties as a group because $r = 0$.

Table 2. Estimates of political web graph size

	n	r	s	$\hat{v}_1$	std. dev. $\hat{v}_1$	$\hat{v}_1/n$
System						
Far Left	16	5	410	1246	548	78
Left	25	1	411	9889	9868	396
Centre	13	1	175	2113	2099	163
Right	28	8	428	1473	513	53
Far Right	20	5	328	1266	558	63
Ecologist	10	5	200	370	158	37
Regionalist	5	1	79	321	310	64
Country						
Austria	12	0	297	na	na	na
France	25	1	377	9073	9052	363
Germany	15	1	193	2717	2703	181
Italy	21	1	514	10301	10278	491
UK	17	1	361	5793	5773	341
Australia	27	15	289	528	131	20
All parties	117	42	2031	5726	873	49

21 Italian parties, only one was directed at another Italian party, and each Italian party made an average of 24 links to entities in the second wave of the snowball sample. This compares with a rather "inward" focus of Australian parties: 15 of the outbound links from Australian parties were to other Australian parties, and each Australian party made an average of only 11 links to entities in the second wave of the snowball sample.

4 CONCLUSIONS

In this paper, an approach for using statistical inference in the analysis of the online networking behavior of political parties has been proposed. The concept of the political web graph has been introduced as has a method for estimating its size using adaptive sampling. Data for 117 political parties from Austria, France, Germany, Italy, UK and Australia were used in the estimation of the sizes of different political web graphs and there is evidence of significant differences in the networking behavior of parties from different party systems and countries. It should be noted that the methods presented here are general and can be applied to the statistical analysis of the

online networking behavior of *any* organization or individual.

The methods and results presented here are preliminary and future work will address the following issues:

- Only one of the estimators discussed in Frank and Snijders (1994) has been presented. There are other estimators in Frank and Snijders (1994) that may be more appropriate.

- As with Frank and Snijders (1994), only the first "wave" of sample data were used in constructing the estimates of web community size - the methods should be extended to include additional waves of data.

- At this stage, no attempt has been made to classify web pages found by the web crawler as "political" in nature or not and it is to be expected that a large proportion of the websites in our snowball reflect topic drift and should be removed. In future research, a website will only be included into the snowball sample if (based on judgement of the researcher) it has been linked to for political reasons. This will improve the estimates of political party web graph size. See Thelwall (2004a)[Chapter 3] for relevant discussion on approaches for assigning value to links between websites.

- A well-defined probability sampling procedure has not been used to obtain the initial samples of political party sites. The initial samples are not Bernoulli samples (but it is not clear that this invalidates comparisons across the party systems) and future research will focus on making adjustments for varying selection probabilities in the initial samples.

- While the links between organizations are modeled as stochastic processes, no allowance for the influence of organization type on link formation is allowed. Following Thompson and Frank (2000), it would be useful to model the formation of links as depending on the characteristics of the organizations and also be dependent on the direction of the link.

REFERENCES

Ackland, R. (2005). `uberlink`: Software for analysing networks on the WWW (user guide). mimeograph, The Australian National University.

Ackland, R. and Gibson, R. (2005). Mapping political party networks on the WWW: How active are the far right? Under review. Available at: `http://acsr.anu.edu.au/staff/ackland/papers/far_right_political_networks.pdf`.

Albert, A., Jeong, H., and Barabasi, A.-L. (1999). Diameter of the World Wide Web. *Nature*, 401:130–131.

Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Brunn, S. and Dodge, M. (2001). Mapping the 'worlds' of the World Wide Web. *American Behavioural Scientist*, 44(10):1717–1739.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10(1):53–67.

Hindman, M., Tsioutsiouliklis, K., and Johnson, J. (2003). 'Googlearchy': How a few heavily-linked sites dominate politics on the Web. mimeograph, Princeton University, 2003. Available at: `http://www.princeton.edu/ mhindman/googlearchy--hindman.pdf`.

Thelwall, M. (2002). Conceptualizing documentation on the web: an evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society of Information Science and Technology*, 53(12):995–1005.

Thelwall, M. (2004a). *Link Analysis: An Information Science Approach*. Academic Press.

Thelwall, M. (2004b). Methods for reporting on the targets of links from national systems of university web sites. *Information Processing & Management*, 40(1):125–144.

Thelwall, M. and Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society of Information Science and Technology*, 54(7):594–602.

Thelwall, M. and Wilkinson, D. (2003). Three target document range metrics for university web sites. *Journal of the American Society of Information Science and Technology*, 54(6):489–496.

Thompson, S. and Collins, L. (2002). Adaptive sampling in research on risk-related behaviours. *Drug and Alcohol Dependence*, 68:S57–S67.

Thompson, S. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26:87–98.

Thompson, S. and Seber, G. (1996). *Adaptive Sampling*. John Wiley & Sons, New York.