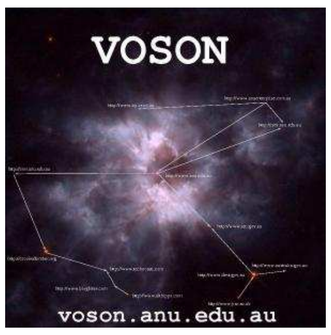


Virtual Observatory for the Study of Online Networks (VOSON)⋆

Administrative organisation: The Australian National University
Collaborating organisations: Oxford University, University of California, Santa Barbara
Chief investigators: Robert Ackland†, Rachel Gibson†, Mathieu O'Neil†, Markus Buchhorn†
Partner investigators: Bruce Bimber‡, Steve Ward§

Aims

The Virtual Observatory for the Study of Online Networks (VOSON)[a] is being established to facilitate innovative and collaborative social science research into the existence and impact of networks on the Internet. The development of VOSON draws upon research methods from both the information and social sciences, incorporating data visualisation, web mining, statistics and more "traditional" empirical social science methods. The research questions underlying the establishment of VOSON require new approaches to data management, computation and resource sharing that involve the use of emerging e-research technologies.

There are two parts to this project. In Part I, a prototype research software tool that will "power" VOSON and can support large-scale international collaborative research into online networks is being developed at the ANU. However, for VOSON to facilitate cutting-edge and collaborative research it is essential to establish international partnerships with researchers who will use the software, contribute to its development, and host VOSON "nodes" at their respective institutions. Part II of the project will focus on the development of these partnerships. In particular, the prototype software environment will be tested in a collaborative "demonstrator" research project involving partner investigators from the University of California, Santa Barbara and at Oxford University.

Development of prototype research software

Under a current 3-year ARC Discovery Grant[b] significant progress has been made toward the development of new methods for the study of online networks formed through hyperlinks. The methods facilitate the visualisation of online networks formed by parties and other political organisations, quantitative analysis of the relative visibility of different types of organisations and characterisation of online political web communities [1].

The research is being conducted using the `uberlink` software for the collection and analysis of online network data (Figure 1). `uberlink` is web-based software that is built using open-source software components and features: PhP/javascript web interface, MySQL database, Perl-based web crawler and interface to the Google API (to allow, for example, the identification of web pages that link to a political party homepage), data manipulation and analysis code in Perl and C++.

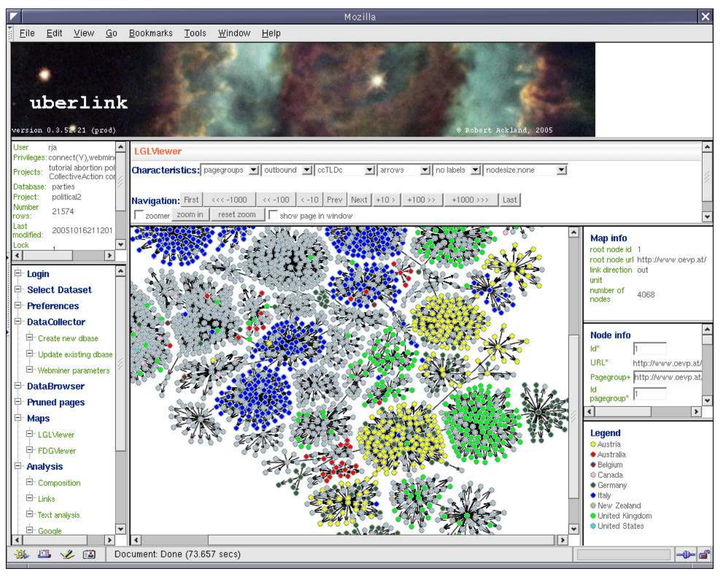


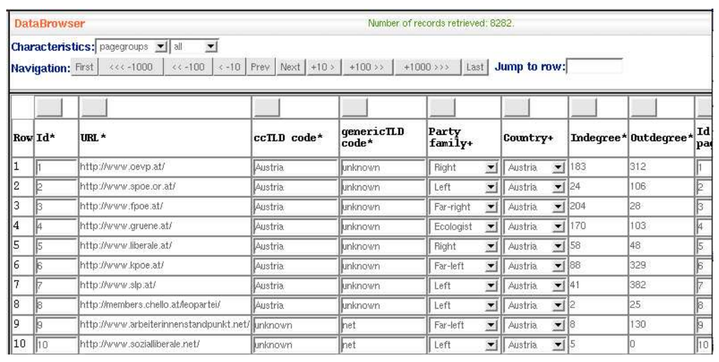
Figure 1: `uberlink` screenshot

Web mining. `uberlink` uses a purpose-built web crawler to return web pages which are then processed to identify hyperlinks between pages. The web crawler can potentially return huge volumes of web data, thereby involving major computational and storage resources. Research into the evolution of online networks requires data collection to occur at regular points in time.

[a] `http://voson.anu.edu.au`

[b] "New Methods for Researching the Existence and Impact of Political Networks on the WWW" (DP0452051), Chief Investigators - Robert Ackland and Rachel Gibson.

Data preparation. Web data are inherently noisy - the inclusion of irrelevant pages ("topic drift") needs to be minimised. Another issue is the unit of analysis - while data are collected at the page-level, analysis is generally conducted over aggregations of pages - a method is required for meaningfully aggregating web pages into pagegroups or "sites".[a] Finally, a method for classifying sites according to the particular domain of study is required. `uberlink` facilitates these three data preparation activities via the DataBrowser (Figure 2). We are also investigating the use of machine learning techniques [specifically, Support Vector Machines - see, for e.g., 7, 4] for categorising pages based on content.



Row	Id*	URL*	ccTLD code*	genericTLD code*	Party family*	Country*	Indegree*	Outdegree*	Id pa
1	1	http://www.oevp.at/	Austria	unknown	Right	Austria	183	312	1
2	2	http://www.spoe.or.at/	Austria	unknown	Left	Austria	24	106	2
3	3	http://www.fpoe.at/	Austria	unknown	Far-right	Austria	204	28	3
4	4	http://www.gruene.at/	Austria	unknown	Ecologist	Austria	170	103	4
5	5	http://www.liberale.at/	Austria	unknown	Right	Austria	58	48	5
6	6	http://www.kpoe.at/	Austria	unknown	Far-left	Austria	88	329	6
7	7	http://www.slp.at/	Austria	unknown	Left	Austria	41	382	7
8	8	http://members.chello.at/leoparteif/	Austria	unknown	Left	Austria	2	25	8
9	9	http://www.arbeiterinnenstandpunkt.net/	unknown	net	Far-left	Austria	8	130	9
10	10	http://www.sozialliberale.net/	unknown	net	Left	Austria	5	0	10

Figure 2: DataBrowser

Data visualisation. We are investigating web-based tools for visualising large network graphs. The directed minimum spanning tree showing all nodes connected to a root node is displayed using the LGL layout algorithm [2](Figure 1). By selecting a node, the characteristics are displayed (and can be edited) in the "node info" panel and thus the LGLViewer is used for both data preparation and analysis.

The LGLViewer provides an abstraction of a network since it shows the shortest path between the root node and all other connected nodes in the database. To visualise all nodes and all links simultaneously we have implemented the LinLog force-directed graphing (FDG) algorithm of [5] which is appropriate for identifying clusters within small-world graphs such as those formed by web data. Web sites are given initial random positions and modelled as electrostatic charges (global repulsion forces). Hyperlinks between web sites are modelled as springs (attraction forces) that move nodes to minimise the energy of the system, thus revealing web clusters or communities. In the FDGViewer (Figure 3) nodes can be selected, edited and zoomed in on. It is possible to clearly see how the formation of perceived clusters in the web map is influenced by the process of data preparation and how the clusters evolve over time.

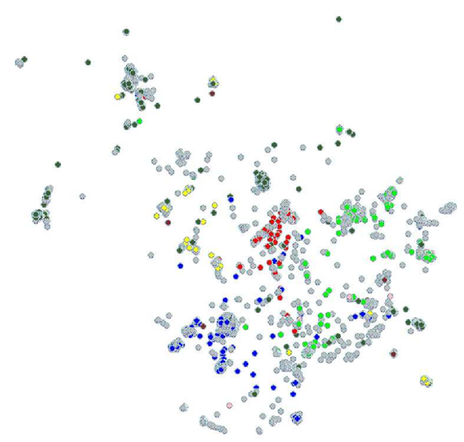


Figure 3: FDG showing web communities

Software - plans. While `uberlink` is currently generating data and analysis for research, there are clear technological constraints relating to data management, computation and resource sharing that prevent large-scale collaborative research. As part of Stage I of this project, we aim to overcome these constraints via the use of e-research technologies, possibly by exposing key features of uberlink (computational/webmining code, visualisation engines, databases) as Grid or web services. The e-Science projects that will be helpful in this goal include: GRENADE (Grid-enabled desktop environment)[b] and RealityGrid (computational steering/distributed visualisation)[c].

[a] See [6] on aggregating pages using "alternative document models" based on directories, domains and multi-domain sites.

[b] `http://www.sve.man.ac.uk/Research/AtoZ/GRENADE`

[c] `http://www.sve.man.ac.uk/Research/AtoZ/RealityGrid`

Collective Action on the WWW

In the demonstrator research we will empirically test a new theory of collective action [Bimber et al., 2006]. The authors have re-conceptualised collective action along two axes: mode of interaction (personal vs. impersonal) and mode of engagement (entrepreneurial vs. institutional), thus producing a four-quadrant "collective action space" which they argue is an appropriate framework for understanding the formation and development of such activities in contemporary society. Bimber et al. (2006) argue that previous models have placed too much emphasis on the decision to participate as a binary phenomenon, and on the formal structures and incentives required to initiate and maintain public engagement. Advances in digital communication technology mean that participation can be more graduated or partial, reducing the level of organisational formality previously required. Creating and disseminating information on the WWW is now almost cost-free, and several types of successful organisations operate in an uncoordinated manner.

The authors place collective action groups (CAGs) such as the recently formed Moveon.org and older organisations like the National Rifle Association (NRA) in one of the four quadrants and provide hypotheses regarding their communication with, and mobilisation of, supporters. Our research will focus on the online presence and use of digital media by CAGs in the US, Australia and UK, providing tests of organisational classification within the four quadrants by examining the structure of web linkages formed by these different types of organisations.

Structural properties of the online networks formed around these groups (i.e. network size, centralisation and inter-connectedness to other networks) will be examined. The relatively democratic nature of the networks will be assessed by testing the hypothesis that linkage patterns found in the "personal and entrepreneurial" quadrant are decentralised, whilst those in the "impersonal and institutional" quadrant are centralised. The project will also assess how varied or homogenous the networks are in terms of international links, and links to other types of organisations and institutions. Finally, an analysis of the evolution of the online "footprint" of CAGs (movement across the collective action space quadrants) will be conducted using data from internet archives such as the Wayback Machine[a].

This research offers an opportunity to test the methods and associated software that are being developed in the VOSON project. It also provides an opportunity to validate and thereby advance an innovative theory regarding technology and the nature of collective action. VOSON will enable these fundamental questions about the changing nature of human behaviour and social organisation to be addressed in different national contexts, thereby strengthening the findings of the research. Ultimately the project is designed to promote a wholly new style of collaborative research that takes advantage of the online environment to produce, share and analyse data not collectable via other means.

References

[1] R. Ackland and R. Gibson. Mapping political party networks on the WWW: How active are the far right? Under review. `http://voson.anu.edu.au/papers/far_right_political_networks.pdf`, 2005.

[2] A. Adai, S. Date, S. Wieland, and E. Marcotte. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. *Journal of Molecular Biology*, 340:179–190, 2004.

[3] B. Bimber, A.J. Flanagin, and C. Stohl. Reconceptualizing collective action in the contemporary media environment. Forthcoming in *Communication Theory*, 2006.

[4] M. Hindman, K. Tsioutsiouliklis, and J.A. Johnson. "Googlearchy": How a few heavily-linked sites dominate politics on the Web. Mimeograph, Princeton University, 2003.

[5] A. Noack. Energy models for drawing clustered small-world graphs. Technical Report 07/03, Institute of Computer Science, Brandenburg University of Technology at Cottbus, 2003.

[6] M. Thelwall. 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, 2002.

[7] V. Vapnick. *The Nature of Statistical Learning*. Springer, New York, 1995.

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

⋆ Australian Research Council (ARC) e-Research Support Project SR0567298. Poster presentation to ARC-DEST Joint Workshop of Funded Projects in *e-Research Support* and *Australian Research Information Infrastructure Committee Schemes*, Monash University, 2-3 November 2005. † The Australian National University, {`robert.ackland,rachel.gibson,mathieu.oneil,markus.buchhorn`}`@anu.edu.au`. ‡ Centre for Information Technology and Society (CITS), University of California, Santa Barbara, `bimber@polsci.ucsb.edu`. § Oxford Internet Institute (OII), Oxford University, `stephen.ward@oii.ox.ac.uk`. This poster was produced using the `a0poster` package for LaTeX.