Virtual Observatory for the Study of Online Networks (VOSON)

Robert Ackland, Rachel Gibson and Mathieu O'Neil

Centre for Social Research, Research School of Social Sciences, The Australian National University†
email: {robert.ackland,rachel.gibson,mathieu.oneil}@anu.edu.au

Introduction

The Virtual Observatory for the Study of Online Networks (VOSON) will facilitate innovative and collaborative research into the existence and impact of social and political networks on the Internet. The development of VOSON draws upon research methods and techniques 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 can only be fulfilled via the use of Grid technology.

Progress to date

Under a current 3-year Australian Research Council Discovery Grant significant progress has been made toward the development of a new methodology, implemented in the research software `uberlink` (Figure 1), to study online networks formed through hyperlinks. Specifically, the work has allowed for 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 political web communities that are being formed online [1, 2, 3].

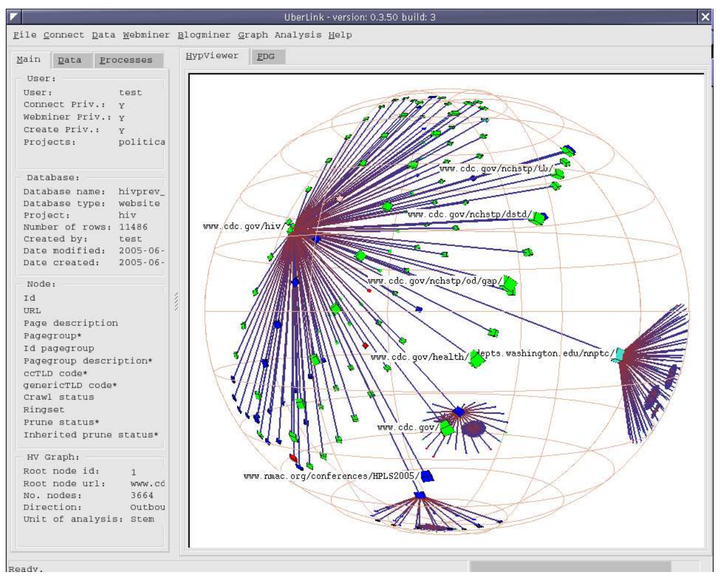

Figure 1: `uberlink` screenshot

`uberlink` has been built using open-source software components. It currently runs on RedHat 8 linux and incorporates a Qt[a] GUI, a MySQL[b] database, a Perl-based web crawler and various data maniupulation and analysis routines programmed in Perl and C++. `uberlink` provides OpenGL visualisation of "cybermaps" using HypViewer [7] graphs and implements a force-directed graphing [6] method for visual identification of web communities (defined as clusters in the web graph).

Grid-enabling `uberlink`

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. We aim to overcome these constraints via the use of Grid technologies, exposing key features of `uberlink` (computational and webmining code, visualisation engines, databases) as Grid services. We are still planning our strategy for Grid-enabling `uberlink`; the next few paragraphs outline some of the issues.

Remote job submissions to HPC resources

One of the key goals of Grid technology is to provide remote access to high-performance computing (HPC) resources (both data storage and computational). Several actions within `uberlink` are computationally intensive.

Web mining. `uberlink` uses a purpose-built web crawler to return web pages which are then processed to identify hyperlinks between pages and page characteristics. 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 - this is not technically feasible in the current environment.

[a] `www.trolltech.com`
[b] `www.mysql.com`

Creation of "pagegroups". The connectivity databases created by `uberlink` contain meta-data pertaining to the web pages - page characteristics and the page's links to other pages within the database. However the analysis is conducted at the *pagegroup* level, where a pagegroup is an aggregation of pages representing an organisational or functional grouping.[a] `uberlink` provides the facility for users to categorise pages and pagegroups in the database and determine the membership of pagegroups. This is achieved via queries to the MySQL server; because the queries may involve hundreds of thousands of web pages, this can be computationally intensive.

Topic drift. One of the challenges with working with web data is dealing with the inclusion of irrelevant pages into the database (e.g. where `www.adobe.com` is picked up by the web crawler because a political party has a link to the Acrobat reader). While `uberlink` has the facility for users to exclude irrelevant pages from the analysis, we are investigating the use of machine learning techniques [specifically, support vector machines - see, for e.g. 5, 11] for categorising pages based on content - again, this is computationally intensive.

GRENADE[b] is a Grid-enabled desktop environment that is relevant to our plans to provide access to HPC resources. GRENADE integrates Globus Toolkit 2.4.3 (GT2)[c] into the open-source K Desktop Environment[d] graphical desktop environment for Linux and Unix workstations.

Computational steering

We are investigating the use of force-directed graphing (FDG) for visual identification of web communities. Web sites (vertices or nodes in the web graph) are given initial random positions and modelled as electrostatic charges (global repulsion forces). Hyperlinks (edges or arcs) between web sites are modelled as springs (attraction forces) that move nodes to minimise the energy of the system. The LinLog FDG algorithm of [8] has been implemented within `uberlink`, and planned research will involve using this algorithm to identify political web communities. With large web graphs, the FDG algorithm is computationally intensive and cannot be viably run on a single workstation. Furthermore, a FDG is a complex system that evolves as it iterates, forms differently according to underlying parameters (e.g. the attractive strength accorded to a hyperlink) and evolves over time as linkage patterns between sites change. We would like to be able to run the FDG algorithm on a remote HPC, but also easily control it, e.g. stop/start, change the parameter space and checkpoint the system for further analysis. We plan to make use of the RealityGrid [e] software to incorporate computational steering into `uberlink`.

Distributed visualisation

`uberlink` provides OpenGL visualisation of the cybermaps and FDGs described above. Small graphs can be adequately rendered on a workstation, but for large graphs we require access to distributed visualisation capabilities. Such visualisation may also be necessary when `uberlink` is demonstrated on the AccessGrid. We will investigate Chromium[f] which provides parallel, distributed OpenGL rendering on commodity clusters.

Current Research into Far-Right Networks

This project will identify the structural properties of the web networks built by the far-right movement in Australia, the United Kingdom and the United States. It will classify the organizational and ideological subgroups of which these networks are composed, as well as analysing their international scope, size and the relative prominence they give to variants of the far-right. This project will also ascertain which organisations or themes receive the greatest amount of common approval in order to bridge movement sectarianism, and analyse connections to outside networks such as mainstream organisations and news outlets.

[a] See [9] 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.globus.org`
[d] `http://www.kde.org`
[e] `http://www.sve.man.ac.uk/Research/AtoZ/RealityGrid`
[f] `http://chromium.sourceforge.net`

We expect far-right movements to be very active users of the Internet given that it represents a "cheaper" form of affiliation than those maintained through other lines of communication and association. The Internet offers particular benefits for those in search of global virtual communities, as it may compensate for their lack of critical mass in their own countries [4]. In addition, stigmatized groups such as the far-right are more likely to use the Internet because of the benefits of confidentiality and decentralisation.

The key stages in the analysis will be: (1) identification of seed sites; (2) data collection, pruning, creation of pagegroups; (3) categorising of sites; and (4) analysis of network structure. We anticipate using adaptive sampling methods [see, for e.g., 10] to help reduce the quantity of web pages and sites that need to be categorised.

Conclusions and Plans

Recently, a web-based version of `uberlink`, `UL-online` has been developed (Figure 2). This provides all of the analytical capabilities of `uberlink` but currently does not support the cybermapping (although we are investigating the use of Java OpenGL[a]).

In addition to the development of prototype Grid-enabled software that will 'power' VOSON, we are focusing on establishing linkages with leading researchers in the UK and the US who are interested in conducting collaborative research into online networks. We contend that the Grid is only as useful as the resources it connects together and VOSON presents a compelling argument in favour of the use of Grid technologies in the social sciences.

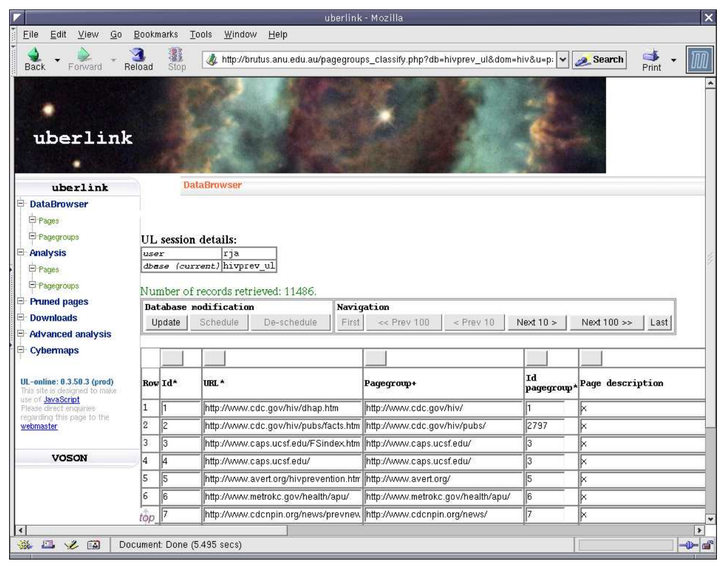

Figure 2: `UL-online` screenshot

References

[1] R. Ackland. Estimating the size of political web graphs. Under review. `http://acsr.anu.edu.au/staff/ackland/papers/political_web_graphs.pdf`, 2005.

[2] R. Ackland and R. Gibson. Mapping political party networks on the WWW. Paper presented at the Australian Electronic Governance Conference, 14-15 April 2004, University of Melbourne. `http://acsr.anu.edu.au/staff/ackland/papers/political_networks.pdf`, 2004.

[3] R. Ackland and R. Gibson. Mapping political party networks on the WWW: How active are the far right? Under review. `http://acsr.anu.edu.au/staff/ackland/papers/far_right_political_network%s.pdf`, 2005.

[4] V. Burris, E. Smith, and A. Strahm. White supremacist networks on the Internet. *Sociological Focus*, 33(2):215–235, 2000.

[5] C. Cortes and V. Vapnick. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[6] P. Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.

[7] T. Munzner. H3: Laying out large directed graphs in 3d hyperbolic space. Proceedings of the 1997 Symposium on Information Visualization, October 2 0-21, 1997.Phoenix, AZ, 1997.

[8] 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.

[9] 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.

[10] S.K. Thompson and G.A.F. Seber. *Adaptive Sampling*. John Wiley & Sons, New York, 1996.

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

[a] `jogl.dev.java.net`

† **Poster presentation at the First International Conference on e-Social Science, 22-24 June 2005, Manchester, UK. This poster was produced using the `a0poster` package for LaTeX.**