The Virtual Observatory for the Study of Online Networks (VOSON) - Progress and Plans

Robert Ackland[1]

[1]Centre for Social Research, Research School of Social Sciences, The Australian National University

Email address of corresponding author: robert.ackland@anu.edu.au

Abstract. The Virtual Observatory for the Study of Online Networks (VOSON) is being established as a Grid-enabled research environment that will facilitate innovative and collaborative research into the existence and impact of social and political networks on the Internet. One of the key software components of VOSON will be the `uberlink` software that has been developed as part of an Australian Research Council-funded project into political networking on the World Wide Web (WWW). While `uberlink` is currently being used for research, there are clear technological constraints relating to data management, computation and resource sharing that prevent its use in large-scale collaborative research. In establishing VOSON, key features of `uberlink` (computational and webmining code, visualisation engines, databases) will be exposed as Grid services, and this paper will describe some of the key planned features of the software environment underlying VOSON. However for VOSON to be successful in facilitating cutting-edge and collaborative research, it is essential to establish international partnerships with researchers who are committed to using the software and collaborating in its further development. The paper will also briefly describe a planned VOSON demonstrator research project involving researchers from Australia, the UK and the US.

Introduction

Over the past 10 years, the Internet has quickly become an important tool for communication, with individuals and organisations increasingly using e-mail and the World Wide Web (WWW) for day-to-day personal and professional interactions. The WWW is particularly effective in facilitating the formation and maintenance of networks and the Web has become vital for networking in business, political and social spheres. The Virtual Observatory for the Study of Online Networks (VOSON) is being designed to provide Grid-enabled research infrastructure to support quantitative social science research into the formation, maintenance and impact of networks on the WWW.

This project is social science e-research (or e-Social Science) in two main senses of the term. First, as stated by Rodden (2004), "the social is increasingly digital" - human activity is increasing becoming manifest in the digital, as well as the physical world, as evidenced by the rise of eSociety, eBusiness, eMedicine and eLearning. The challenge for social scientists is to develop new methods for collecting and analysing digital traces of human activity and the development of these methods will draw from fields outside of the social sciences, such as the computer and information sciences. VOSON involves the development of methods and associated software for the collection and analysis of Web data that can provide new insights into the online networking activities of individuals and organisations.

Second, VOSON involves the use of emerging Grid technologies that are the hallmark of eScience. The WWW is a vast and growing universe of digital data and Web research involves the creation and analysis of potentially huge datasets. Any project involving the use of large-scale datasets will benefit from (or indeed, may depend on) the use of Grid technologies for data storage and access to computational resources on high-performance computers (HPC). Grid technology will also help to facilitate collaboration amongst researchers in different locations (a planned VOSON project involves political scientists in Australia, the US and the UK working on comparative research into networking on the Web by different social organisations). For such collaborative research to succeed, it will be necessary for researchers to have (often simultaneous) access to shared datasets and computational resources – something that the Grid is being designed to facilitate.

Research into social and political networks on the Internet

It has been argued (see, for example, Castells, 1996) that while social networks have always existed, it was only when information and communications technology (ICT) became ubiquitous that networks could fully flourish. Exploration of the existence and impact of ICT-enabled networks has formed a significant focus of political and social science research into the Internet. However social scientists currently do not have access to appropriate research infrastructure for the efficient collection and analysis of Web data. Web data is vastly different both in format and volume to the (survey-based) data usually studied by empirical social scientists, and it is necessary to look to the research of computer and information scientists for insights into the appropriate tools and methodologies. Under a current 3-year Australian Research Council Discovery Grant, significant progress has been made in the development of a new methodology, implemented in the research software `uberlink` (Figure 1), to study online networks formed through hyperlinks. Specifically, the research project is focusing on visualisation of online networks formed (via hyperlinks) by political parties and other organisations, quantitative analysis of the relative visibility of different types of organisations and characterisation of web communities that are being formed online (Ackland, 2005a, 2005b; Ackland and Gibson, 2004, 2005; Ackland and Gray, 2005).

One of the main themes to emerge from this research is whether it is possible to identify significant differences in the online networking behaviour of socially or politically conservative-leaning organisations, compared with their counterparts on the left, and if so, what are the implications of such differences in networking behaviour? There is evidence to suggest that conservative-leaning organisations or groups tend to be more active in their linking behaviour and form more dense online communities. Adamic and Glance (2005) and Ackland (2005b) show that conservative "A-list" political webloggers form more dense patterns of linkages compared with prominent liberal A-list bloggers. Adamic (1999), in a study of the web behaviour of organisations involved in the abortion debate, similarly found that pro-life groups formed denser online networks, compared with their pro-choice counterparts. The reason why differences in the online networking behaviour of right versus left organisations is important is that it has implications for the relative visibility of different political messages and ideologies on the WWW. As discussed in Hindman et al. (2003), visibility on the web is a relative concept, and the visibility of a given web site is largely influenced by the number of inbound links to that site (Google, for example, ranks more

heavily-inlinked sites higher, because the number of inbound links to a site is seen as a measure of “authority” on a particular topic).

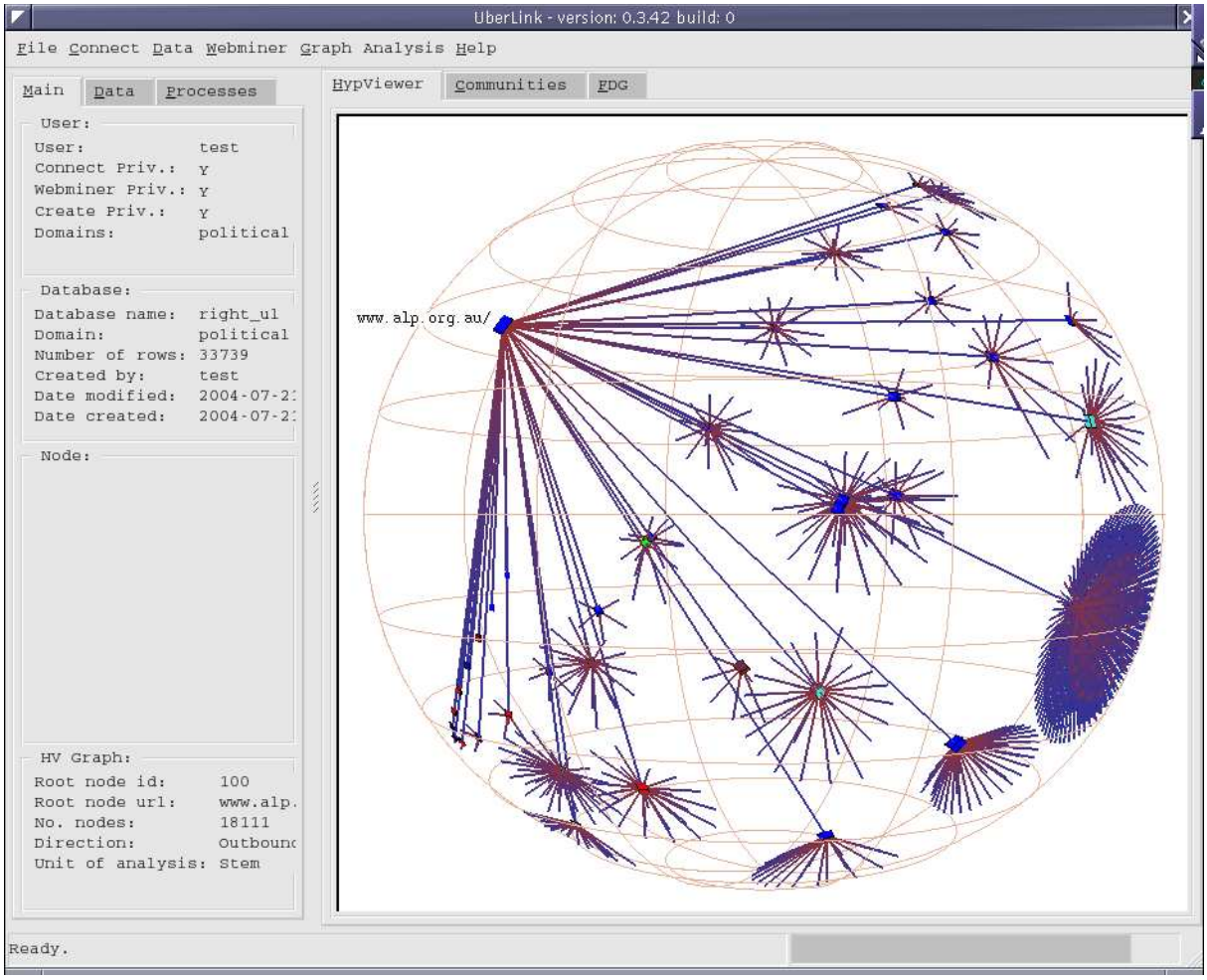

Figure 1: `uberlink` screenshot

Grid-based computing and data management

Grid technology is still evolving: at present, there are no simple ‘off-the-shelf’ solutions for easily Grid-enabling a research project. Moreover, Chin and Coveney (2004) argue that significant barriers to widespread acceptance of the Grid by application scientists remain, such as large and cumbersome toolkits, significant administrative overhead associated with configuration, poor documentation, inappropriate focus on creating middleware bindings in Java rather than C/C++/Fortran, and lack of facility for transferring binary data across the Grid. The authors conclude somewhat depressingly that “...the state of the Grid is still such that there are no significant benefits offered to application scientists, who currently consider it easier (and more reliable) to submit and control tasks through more traditional means such as SSH.” (Chin and Coveney, 2004, page 5). Social scientists, however, are unlikely to use these traditional means, and hence are more reliant than application scientists on effective and relatively easy-to-use Grid solutions. Without a tradition of scientific computing they face a steep learning curve. One UK-based social scientist even commented on the ‘heroic’ effort required effort to use Grid technology (Crouchley, 2004). In this context, our objective is not

to develop production code, but to provide a software testbed that will help us to evaluate Grid technologies in the context of our particular research needs.

Some relevant Grid projects

The GRENADE[1] project (Foster et al., 2004; Pickles et al., 2004) has released a prototype Grid-enabled desktop environment and is encouraging end-users and developers to use and extend this software. GRENADE integrates Globus Toolkit 2.4.3 (GT2)[2] into the open source K Desktop Environment[3] graphical desktop environment for Linux and Unix workstations. The software provides Konqueror (the KDE file browser) plug-ins for using the GRIP protocol to query available Grid resources for status information, plug-ins for browsing remote file systems, and GUIs (using the Qt[4] C++ GUI Toolkit) for job definition, submission and management. The GRENADE team aims to make Grid computing ubiquitous by bringing it to the desktop and a stated aim is for more "collaborative use of the Grid to become commonplace ... [via] a more intuitive and friendly means of interaction" (Foster et al., 2004, page 1).

RealityGrid[5] (Chin et al., 2003; Chin, 2003; Chin and Coveney, 2004; Coveney et al., 2004) is a collection of software for making research applications steerable in the sense that the scientist can remotely interact with a high-performance computer (HPC)-hosted application (e.g. a simulation), monitoring and altering system parameters and creating 'checkpoints' or snapshots of the system to disk. RealityGrid (ReG) has focused on Grid-enabling two pre-existing applications for simulations of complex fluids in two and three dimensions (LB2D and LB3D). LB3D was Grid-enabled using the ReG steering library, which in turn used GT2 for resource discovery and communication. However, GT2 was found to be impractical to work with - even file transfer was difficult because of problems with firewalls and the inability to copy multiple files in a single invocation. The Grid-enabling of LB2D was instead achieved using an alternative middleware design - `OGSI::Lite` (McKeown, 2005) which is a Perl-based container to allow simple creation of Grid Services. The ReG team were able to quickly build a simple remote job submission and steering system using LB2D and OGSI::Lite and provided a Qt-based GUI steering client. The ReG team recently introduced WEDS, a WSRF-based middleware scheme that is proposed as an alternative to Globus and an improvement over `OGSI::Lite`.

Relevant UK e-Social Science projects

The INWA project[6] has developed a Grid-enabled corporate data mining environment for consumer demand modeling and management of customer relationships, in the telecommunications, financial and property sectors. Data mining is the use of automated procedures in the discovery of information from large quantities of data. INWA enables analysts to remotely and securely submit data mining batch jobs that are run on a HPC environment local to the data, with the results then transferred back to the analyst. INWA uses GT2 for the Grid middleware, Grid Engine[7] as the compute resource manager, Transfer-

1 `http://www.sve.man.ac.uk/Research/AtoZ/GRENADE`
2 `http://www.globus.org`
3 `http://www.kde.org`
4 `http://www.trolltech.com`
5 `http://www.sve.man.ac.uk/Research/AtoZ/RealityGrid`
6 `http://www.epcc.ed.ac.uk/projects/inwa/`
7 `http://gridengine.sunsource.net`

queue Over Globus[8] to transfer batch jobs and results between local and remote sites, and OGSA-DAI[9] for access to relational databases via the Grid.

The SABRE/R project (Crouchley, 2004) aims to demonstrate the effectiveness of an OGSA component-based approach to middleware for handling complex statistical modeling problems. SABRE[10] provides for the statistical analysis of binary, ordinal and count recurrent events, and R[11] is a language and environment for statistics and graphics. Unlike INWA, the SABRE/R project is an application of Grid technologies to 'traditional' empirical social science research using relatively small but highly-complex longitudinal datasets such as the British Household Panel Survey. As discussed in Crouchley (2004), social science statistical analysis of longitudinal data can involve very complex modeling to deal with, among other things: cluster effects, measurement error, missing data/dropout/sample selection, and endogeneity. This complexity leads to heavy computational requirements and the benefits of a computational Grid are evident.

VOSON will involve webmining (the application of data mining to the WWW) of very large quantities of data, thus requiring access to both data and computational Grids. Both VOSON and the SABRE/R project are concerned with the use of Grid technologies to answer core social science issues studied by economists, sociologists and political scientists. The difference is that VOSON is focused on the study of how these processes are carried out in the online world, as well as the impact of the Internet on 'offline' social and political phenomena. While SABRE/R does not involve the collection of new data nor the development of new methods of analysis, VOSON will use Grid technology for the collection and analysis of new forms of social science data from the Internet.

Grid-enabling `uberlink`

While `uberlink` already runs on a workstation and is 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. VOSON involves the investigation of methods for overcoming these constraints via the use of Grid technologies. Our aim is to expose key features of `uberlink` (computational and webmining code, visualisation engines, databases) as Grid services, thus creating a prototype software environment made of various tools that can facilitate international collaborative research into online networks. GRENADE will serve as the basis for resource discovery, data management and job submission. The familiar desktop environment facilitated by GRENADE is key to our plans to provide a collaborative research environment that social scientists are comfortable with. Computational steering will be provided by the ReG software.

Hardware components in the Grid testbed

All hardware in our Grid testbed environment will run a variant of RedHat Linux[12] (e.g. Fedora Core 3). The proposed setup is as follows (see Figure 2):

- local workstation (`local`). A machine an analyst works from.

[8] `http://gridengine.sunsource.net/project/gridengine/tog.html`
[9] `http://www.ogsa-dai.org`
[10] `http://www.cas.lancs.ac.uk/software/sabre3.1/sabre.html`
[11] `http://www.r-project.org`
[12] `http://www.redhat.com`

• remote ‘head’ server (`head`). This machine receives remote job submissions and computational steering requests, farms these requests out to `compute` and sends the results back to `local`.

• remote compute HPC cluster (`compute`). An openMosix[13] cluster on which standard Linux applications/processes can be run without the need for rewriting the code to access MPI libraries (unlike Beowulf clusters, for example).

• remote visualisation HPC cluster (`viz`). Chromium[14] provides parallel, distributed OpenGL rendering on commodity clusters.

• remote MySQL[15] server (`dataServe`).

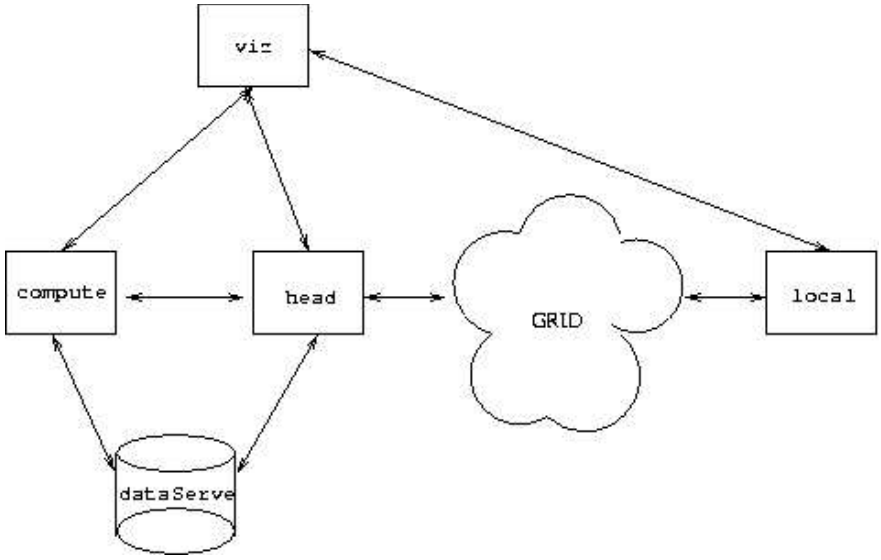

Figure 2: VOSON Grid testbed

Remote job submissions to HPC resources

One of the key goals of Grid technology is to provide remote access to HPC resources (both data storage and computational). There are several actions within `uberlink` that are computationally intensive; here we focus on two such actions.

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. Even the relatively small-scale research conducted to date involves the web crawler running continuously for days at a time (significantly reducing the performance of the workstation running `uberlink`) and the collection of gigabytes of web pages (which are currently not stored because of lack of capacity). 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.

[13] `http://openmosix.sourceforge.net`
[14] `http://chromium.sourceforge.net`
[15] `http://www.mysql.com`

Changes to pages in connectivity databases. 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 *page group* level, where a page group is an aggregation of pages representing an organisational or functional grouping (see Thelwall, 2002 on aggregating pages using 'alternative document models' based on directories, domains and multi-domain sites). `uberlink` provides the facility for manually changing characteristics of pages in the database. For example, instances of *topic drift* (e.g. where `www.adobe.com` is picked up by the web crawler because a political party has a link to the acrobat reader) are dealt with by the analyst flagging the page (and associated page group) as irrelevant, and therefore to be excluded from further analysis and web crawler runs. This is achieved via a query to the MySQL server; but because there can be thousands of pages in a given page group (and many page groups that need to be changed in a given query), this is computationally intensive.

Grid strategy for remote job submission. GT2 will run on `head`, while GRENADE, GT2 and `uberlink` will run on `local`. For the prototype, `local` will also run a MySQL server and the connectivity databases will be stored both on `local` and `dataServe` (synchronisation of the databases will be achieved via a method described below). GRENADE will be used to handle sign-on and resource discovery. In `uberlink`, the user will initiate queries to create/modify page groups and requests to start/stop the web crawler. These jobs will be submitted to `head` via the GRENADE interface, and this interface will also be used to inform the user of the status of remotely executed jobs.

Data management

While it is feasible for different researchers to simultaneously work on the same dataset using methods proposed in (a) above, there is a potential problem regarding versioning of databases that can be solved by Grid technology.

Grid strategy for data management. GRENADE can periodically poll `head` to test whether the local version of the database is up-to-date. If the master version stored on `dataServe` is more current, then the user will be alerted and prompted to download the most recent version using the GridFTP utility in GT2. GRENADE will also be applied to ensure that multiple users do not try to simultaneously submit commands involving the same database.

Computational steering

It is possible to use force-directed graphing (FDG) (Eades, 1984) for 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 Noack (2003) 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.

Grid strategy for computational steering. The version of the ReG software that uses `OGSI::Lite` and a Qt steering client has been tested, and can be used for incorporating computational steering into `uberlink`. Via GRENADE, the analyst will launch a steering-enabled version of the FDG code present in `uberlink` which will run on `compute`. `uberlink` will be modified to incorporate code from the ReG Qt-based steering client, thus facilitating remote steering of the FDG algorithm. `uberlink` currently provides an OpenGL visualisation of the FDG - the input to this visualisation is a text file containing the website coordinates (in 3D space). At each step of the FDG algorithm, a new text file will be checkpointed and transmitted to `local` via GridFTP so the analyst can locally view the FDG system as it iterates. The Grid strategy will thus involve using software from both the ReG and GRENADE projects. We note that the GRENADE team anticipate that "the RealityGrid project will take advantage of the GRENADE software"[16], and are therefore confident that our Grid strategy is sound, and will be informed by further developments within the ReG and GRENADE projects over the lifetime of this project.

Distributed visualisation

`uberlink` currently provides OpenGL visualisation of HypViewer (Munzner, 1997) graphs and the 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 will also be necessary when `uberlink` is demonstrated on tiled computer monitors (e.g. the AccessGrid).

Grid strategy for distributed visualisation. We will investigate the use of the visualisation server (`viz`) described above. We note that the GRENADE team have flagged integration of SGI's OpenGL VizServer[17] as a possible further extension, and such extensions will benefit the development of distributed visualisation in VOSON.

Conclusions

This paper has described progress and plans for the establishment of VOSON, a Grid-enabled research environment that will facilitate international collaborative research into the existence and impact of networks on the Internet. In addition to the development of prototype Grid-enabled software that will ' powerVOSON, it is obviously essential to establish international linkages with researchers who are interested accessing VOSON data and analytical capabilities, and being involved in the future development of the project. To this end, a VOSON-facilitated collaborative demonstrator project involving researchers from the Australian National University, the Oxford University and the University of California, Santa Barbara is being planned. The aim of this research project is to conduct a cross-country comparison of how old and new social movements are using the Internet to network with other groups and mobilise support, and the extent to which technology lowers the cost of collective action.

---

[16] GRENADE proposal document (p. 4). Available at: `http://cascade.man.ac.uk/uploads/1045737822-resources-155.pdf`. Accessed on 3rd May, 2005.

[17] `http://www.sgi.com/products/software/vizserver/`

References

Ackland, R. (2005a). Estimating the size of political web graphs. Under review. Available at: http://acsr.anu.edu.au/staff/ackland/papers/political_web_graphs.pdf. Accessed: 3rd May, 2005.

Ackland, R. (2005b). Mapping the US political blogosphere: Are conservative bloggers more prominent? Presentation to BlogTalk Downunder 2005, 19-21 May 2005, Sydney. Available at: http://acsr.anu.edu.au/staff/ackland/papers/polblogs.pdf. Accessed: 3rd May, 2005.

Ackland, R. and Gibson, R. (2004). Mapping political party networks on the WWW. Paper presented at the Australian Electronic Governance Conference, 14-15 April 2004, University of Melbourne. Available at: http://acsr.anu.edu.au/staff/ackland/papers/political_networks.pdf. Accessed: 3rd May, 2005.

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. Accessed: 3rd May, 2005.

Ackland, R. and Gray, E. (2005). Australia's online presence as encountered by potential migrants. Under review. Available at: http://acsr.anu.edu.au/staff/ackland/papers/online_presence.pdf. Accessed: 3rd May, 2005.

Adamic, L. (1999). The small world web. In *Proceedings of the 3rd European Conf. on Digital Libraries*, volume 1696 of *Lecture notes in Computer Science*, pages 443-452. Springer, 1999. Available at: http://www.hpl.hp.com/shl/papers/smallworld/smallworldpaper.html. Accessed: 3rd May, 2005.

Adamic, L. and Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. Mimeograph. Available at: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf. Accessed: 3rd May, 2005.

Castells, M. (1996). *The rise of the network society. The information age: Economy, society and culture Vol. I*. Blackwell, London.

Chin, J. (2003). Adventures with LB2D and OGSI::Lite. Availabe at: http://the.earth.li/~jon/work/ogsi/writeup/. Accessed: 3rd May, 2005.

Chin, J. and Coveney, P. (2004). Towards tractable toolkits for the grid: A plea for lightweight, usable middleware. UK e-Science Technical Report UKeS-2004-01. Available at: http://www.nesc.ac.uk/technical_papers/UKeS-2004-01.pdf. Accessed: 3rd May, 2005.

Chin, J., Harting, J., Jha, S., Coveney, P., Porter, A., and Pickles, S. (2003). Steering in computational science: mesoscale modelling and simulation. *Contemporary Physics*, 44:417–434.

Coveney, P., Vicary, J., Chin, J., and Harvey, M. (2004). Introducing WEDS: A WSRF-based environment for distributed simulation. UK e-Science Technical Report UKeS-2004-07. Available at: http://www.nesc.ac.uk/technical_papers/UKeS-2004-07.pdf. Accessed: 3rd May, 2005.

Crouchley, R. (2004). An OGSA component-based approach to middleware for statistical modelling and introduction to CQeSSS. National Centre for e-Social Science All Hands Meeting, 5-6 July 2004. Available at: http://tyne.dl.ac.uk/ReDRESS/CQeSSR2.ppt. Accessed: 3rd May, 2005.

Eades, P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160.

Foster, M., Hanlon, D., MacLaren, J., Marsh, J., Pettifer, S., and Pickles, S. (2004). Grid-enabled desktop environments: The GRENADE project. Paper delivered at UK e-Science All Hands Meeting, Nottingham, September 2004. Available at: http://www.allhands.org.uk/proceedings/papers/291.pdf. Accessed: 3rd May, 2005.

Hindman, M., K. Tsioutsiouliklis, and J. Johnson (2003): ' Googlearchy'How a Few Heavily-Linked Sites Dominate Politics on the Web. Princeton University, 2003. Available at: http://www.princeton.edu/~mhindman/googlearchy--hindman.pdf. Accessed: 3rd May, 2005.

McKeown, M. (2005). WSRF::Lite and OGSI::Lite. Available at: http://www.sve.man.ac.uk/Research/AtoZ/ILCT. Accessed: 3rd May, 2005.

Munzner, T. (1997). 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. Available at: http://graphics.stanford.edu/papers/h3/. Accessed: 3rd May, 2005.

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

Pickles, S., Foster, M., MacLaren, J., Marsh, J., and Pettifer, S. (2004). Grid-enabled desktop environments: The GRENADE project. Paper delivered at UK e-Science All Hands Meeting, Nottingham, September 2003. Available at: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/134.pdf. Accessed: 3rd May, 2005.

Rodden, T. (2004), What is e-Social Science? An introduction to the ESRC e-Social Science Strategy, NCeSS All Hands Meeting, Manchester, 5-6 July 2004. Available at: http://www.ncess.ac.uk/conference-05/past/ncess_ahm_rodden.pdf. Accessed: 3rd May, 2005.

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.