Using Facebook as a Data Source and Platform for e-Researching Social Networks

Robert Ackland[1]

[1]Australian Demographic and Social Research Institute, The Australian National University

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

Abstract. Social networking services (SNS) such as Facebook and Orkut enable new research into the role of individual characteristics in friendship patterns and the diffusion of tastes in social networks. This paper assesses the opportunities and challenges posed by SNSs for empirical research into online social networks. It is argued that SNSs may eventually provide platforms for delivering social network analysis e-Research tools, and a prototype tool built using the OpenSocial API is presented. MyExperiment has been described as "Facebook for scientists"; this paper contends that that SNSs such as Facebook may eventually be described as "GridSphere for e-Social Scientists".

Introduction

There are two types of networks on the Web: networks of websites or blogsites (or individual pages therein) where the network tie is the hyperlink, and networks of individuals in social networking services (SNS) such as Facebook, where the ties are "friendships" (where a user requests and gains permission to list another user as a friend on his or her profile) or joint membership in groups (for example, for alumni of particular universities). This paper first provides a summary of empirical research into online networks, comparing hyperlink analysis with analysis of social networks enabled by SNSs. The potential role of SNSs as platforms for delivering e-Research tools to social scientists is then discussed and VOSON-os, a prototype social network analysis tool delivered on the OpenSocial platform, is presented.

Empirical research into online networks

Hyperlink networks

There are three broad approaches for empirical research into hyperlink networks, each reflecting a particular disciplinary base.[1] First, "webmetrics" (also known as webometrics and cybermetrics) is an approach for analysing hyperlink data and website usage patterns that largely draws on bibliometrics and infometrics and is most commonly used in the library and informtion sciences (LIS) - see, for example, Thelwall, Vaughan, and Björneborn (2005). A recent application of webmetrics (Barjak and Thelwall, forthcoming) involves regressing counts of inbound hyperlinks to the websites of life science research teams on relevant offline characteristics of the teams (e.g. gender of team leader, industry connections, research productivity) in order to assess the role of hyperlinks as science and technology indicators.

1 Note that there is in fact active cross-over between the various approaches.

Second, applied physicists have focused on the identification of empirical properties in large-scale collections of Web pages and the development of statistical-mechanical models that can be used to explain these properties.[2] For example, Newman (2002) studied the existence of "assortative mixing" (or correlation between the attributes of adjacent network nodes) on the Web. Barabási and Albert (1999) explain "power laws" in the distribution of links in networks such as the Web, where a small number of sites receive the lion's share of links pointing toward them, via the concept of preferential attachment: newer entrants are inclined to link to already well-connected actors, thereby increasing the incumbents' advantage.

Finally, social science approaches to hyperlink analysis have attempted to understand the role of the Web in enabling various forms of social, economic or political behaviour. An obvious aspect that distinguishes social science hyperlink analysis is a focus on actors traditionally studied by social scientists, for example, political parties (Ackland and Gibson, 2004), environmental social movement organisations (Ackland, O'Neil, Bimber, Gibson, and Ward, 2006), and civil society actors (González-Bailó, 2007). Social science hyperlink analysis also involves testing of hypotheses emerging from social science models. For example, Shumate and Dewitt (2008) identify "structural signatures" of collective action behaviour in hyperlink networks, whereby HIV/AIDS NGOs are creating an "information public good". Compared with researchers from LIS and applied physics, social scientists are more inclined to view the Web as a relational space and this leads to social network analysis (SNA) being used to analyse how the structural position of actors impacts on actions and opportunities.

It has taken longer for empirical hyperlink analysis to develop and gain popularity within the social sciences, compared with LIS and applied physics, for three main reasons. The first relates to the conceptual framework for studying the Web, and the theoretical meaning of hyperlinks and network nodes. An applied physicist's view of the Web as a large-scale network of hyperlinked documents is non-controversial in that this is the underlying architecture of the Web. Applied physicists are generally not concerned with attributing theoretical or behavioural meaning to hyperlinks, and focus rather on developing sophisticated approaches for characterising and simulating hyperlink networks. Researchers from LIS are more likely (compared with applied physicists) to be interested in the behavioural foundations of hyperlinks. However, in contrast with social scientists, LIS researchers have more readily embraced hyperlinks as having theoretical meaning, possibly because they can be seen to be analogous to citation networks traditionally studied within this field. To the extent that LIS researchers study scholarly activity via hyperlink analysis, the concept of network node is more complicated than the individual web pages studied by applied physicists, but there are well-established LIS approaches (known as "alternative document models") for aggregating web pages up to appropriate units of analysis e.g. research teams, universities. Social scientists view the Web not as a network of documents but as a network of people and organisations (a view that is more closely aligned with that of LIS researchers). This world view leads to greater difficulty in developing conceptual frameworks for hyperlink networks; in the context of research into online social movements, for example, what does a hyperlink mean, and what are the appropriate network nodes?

The second reason why social scientists have been slower to embrace hyperlink analysis relates to research methods, and in particular, the challenge of dealing with datasets that can potentially include millions of web pages. The scale of such data does not pose a problem for applied physicists since they rarely require information on network nodes and ties other than that that can be obtained via automatic means (in particular, there is a focus on network attributes that are derived from the graph structure itself, e.g. indegree or outdegree, rather

2 This work has been developed in the context of large-scale networks in general, with the Web as one such example.

than attributes that pertain to a web page, web site or the owner of the website). While LIS researchers are interested in obtaining attribute data for websites representing, for example, research teams, webmetrics does not deal with complete networks. It focuses exclusively on local- or ego-networks (e.g. finding, for a sample of websites, a total count of hyperlinks pointing to the website and using this as a measure of scholarly visibility), and there is no need to attach attributes to the *senders* of these hyperlinks (the focus is rather on the attributes of the *receivers* of hyperlinks). Social science hyperlink research (in particular, where SNA is used) involves complete networks, so you either need a small and well-defined population of sites obtained from a sampling frame that is external to the network (e.g Ackland and Gibson, 2004, study political party hyperlink networks) or else you need an approach for constructing network samples. Sampling is a cornerstone of empirical social science research but it is not currently used in hyperlink analysis. In fact, there is some resistance from computer scientists to the idea that limitations of automated data collection necessitate network sampling in Web research (this goes against the Semantic Web, defined as "the web of data with meaning in the sense that a computer program can learn enough about what it means to process it"[3]).

A final reason why hyperlink analysis has developed more slowly in the social sciences is a lack of appropriate tools. Tool development is better accepted as a part of scientific process in applied physics and LIS (and researchers often have the skills to develop their own tools). For social science research into the Web there needs to be access to a set of heterogeneous tools that can enable observation and tagging of hyperlinks and/or network actors, and aggregation of pages into meaningful groups. While these tools might be adopted from other approaches, they need to be modified to support a social science view of the Web. It is unreasonable to expect that a tool that has been developed for an entirely different research context will make a big impact on social science Web research without major modification. A "Google for Social Scientists", with focus on full-text search over Web pages as the core research function, and no regard for sampling and statistical inference, is unlikely to have a big impact.

Online social networks

We define an online social network here as a network extracted from a SNS such as Facebook or Orkut. Research into online social networks is only just beginning and it is therefore difficult to assess the extent to which social scientists will embrace this new area (compared with reserach into hyperlink networks). However, social scientists appear to be well placed to engage with and undertake analysis of online social networks. Sociologists have been studying friendship networks for decades and have found that patterns of friendship are strongly affected by characteristics such as age, race and language ("birds of a feather flock together"). Online social networks offer new data for research into homophily (see, for example, Wimmer and Lewis, 2008) and also research into the diffusion of tastes.

To what extent are the three constraints identified with respect to social science research into hyperlink networks (difficulty of developing conceptual frameworks, methodological issues, availability of appropriate tools) likely to be different with online social networks? It is arguably easier to develop a conceptual framework for social science research into online social networks. With SNSs, the nodes are clearly people and the network ties are also relatively easily interpreted. While it is obvious that a friendship in Facebook means something different to an offline friendship (in terms of the cost of making and maintaining the tie, for example, and in terms of the public-private nature of the action), it is perhaps easier to interpret a tie in Facebook compared with a hyperlink tie.

3 http://www.w3.org/People/Berners-Lee/Glossary.html#Semantic

The relatively easy interpretation of network nodes and ties also means that social science research into online social networks is possibly less methodologically challenging, compared with hyperlink networks. Users of SNSs are encouraged to describe themselves (e.g. political persuasion, religion, humour, smoking and drinking behaviour) using a combination of text fields, drop-down selection boxes and check boxes. The profiles are highly amenable to automated data analysis (compared with web pages which are much less structured). However, despite the machine-readability of SNSs, there are two reasons to expect that network sampling will still be important for social science research into online social networks. First, SNA techniques such as exponential random graph modelling currently have difficulty with large-scale networks. Further, these techniques are not well-suited to networks where a person can have 5000 “friends”, and there will be a need for further data interrogation (perhaps viewing of the profiles to establish friendship using a different metric or else by directly asking the user to identify their “real” friends). Also, it is possible that individual attribute data needed by social scientists will not be contained on publicly-viewable user profiles (either because it isn’t of interest to the creator of the SNS or else it is of a personal nature), and hence will need to be obtained via follow-up surveys.

Social networking service applications and e-Research

It was argued above that a limiting factor for social science hyperlink research is the availability of appropriate tools. It is proposed here that SNSs can, in addition to providing new data for social science research, serve as delivery platforms for e-Research tools that are specifically catered to social scientists studying online social networks.

VOSON-os - a demonstrator OpenSocial e-Research tool

In order to assess the viability of SNSs as platforms for e-Research tools, a prototype of the Virtual Observatory for the Study of Online Networks (VOSON) e-Research tool[4] was developed using the OpenSocial API. OpenSocial has been developed as a direct challenge to Facebook (who pioneered applications for SNSs).[5] The technical details of developing an OpenSocial application are not covered here (detailed tutorials are available online), but the VOSON-os prototype (Figure 1) was very easy to build (especially when compared with the challenges of working with portal development frameworks such as Gridsphere).

VOSON-os features an interactive network map which is provided using JSViz.[6] The intention is that this map would show the ego network of the person who has installed VOSON-os (an ego network consists of a focal node ("ego") and the nodes directly connected to ego ("alters") plus any ties among the alters).[7] The OpenSocial API enables the programmatic identification of the friends of the user who has loaded the VOSON-os, but in order to know the ties amongst the friends of the user, it is necessary that those users also install VOSON-os. So, we can imagine a research strategy that might involve emailing people in a particular target group and asking them to install VOSON-os, and this would enable the construction of a complete network containing all users who had loaded the application (there would of course be a response rate less than 100%, but this is the same with SNA conducted offline). The application could also be used to elicit further information from the participants

4 http://voson.anu.edu.au

5 http://code.google.com/apis/opensocial/. As indicated by the title of this paper, the original intention was to build a Facebook application, but OpenSocial appeared easier to work with and the potential for using it on different SNSs (it will work under Ning, Orkut, MySpace and a number of other SNSs) was attractive.

6 http://www.jsviz.org/blog/

7 Note that the map in Figure 1 contains dummy data since the author doesn’t actually have any friends on Orkut.

e.g. users could be periodically asked to provide information on labour market status, which could then be correlated with professional networking behaviour.

VOSON-os also features basic graph-level SNA metrics calculated for the dummy data (note: these should actually be calculated for complete networks, rather than an ego network). A Facebook or OpenSocial e-Research SNA tool would preferably make use of SNA libraries such as the sna library for the R statistical software (http://erzuli.ss.uci.edu/R.stuff/) or Jung (http://jung.sourceforge.net/). Ideally, these routines could be accessed via web services (however, the cross-domain security model currently in OpenSocial prevents this).

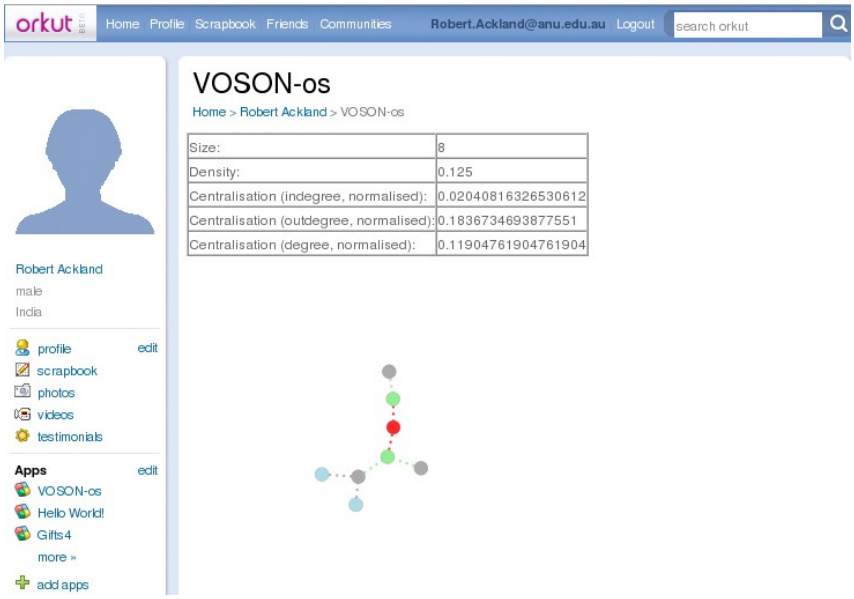

Figure 1. VOSON-os - a demonstrator SNA tool on the OpenSocial platform

The promise of e-Research for social science

Social scientists have been analysing surveys of individuals and households for decades and are well served by existing tools such as Stata, SAS and SPSS. It is in new areas of research, such as research into online networks, that e-Research can make a major contribution to social science, by enabling access to new forms of data and research methods. Social science research into the Web requires diverse tools (e.g. web crawler, text mining, data visualisation, social network analysis) that are unlikely to be provided by a single tool developer (indeed, such "vertical integration" could potentially lead to anti-competitive behaviour).

e-Research promises the seamless connection (via web services) of research resources (data, methods, computational) from different providers. However, there have been problems in how this has been implemented. First, there has been a reliance on heavyweight middleware software such as Globus and portlet development frameworks such as Gridsphere (with the emphasis on JSR-168 compliance and user interfaces that are more suitable to computational scientists rather than social scientists), eschewing simpler, but effective web technologies such as AJAX-enabled websites. Second, the central involvement of technologists in e-Social Science (almost an inevitable consequence of the use of sophisticated technologies such as Globus and Gridsphere) has led to indirect pressure on social scientists to adopt research approaches from other disciplines, regardless of the fact that social scientists already have well-developed and sophisticated research frameworks. The result has been an emphasis on searching over digital research collections (something more important to the humanities than

the social sciences), workflows and ontologies, while important aspects of empirical social science (e.g. sampling and statistical inference) have been relatively neglected.[8]

There needs to be balance between the aims of technologists who must push technology boundaries to receive kudos and the needs of social scientists, who are often just looking for something "quick and dirty" that will enable them to decide whether to invest further in e-Social Science. A social scientist will gain kudos from publications in appropriate journals and the development of research tools is a means to an end, rather than the end in itself.

It has been argued here that SNSs such as Facebook present interesting possibilities as platforms for e-Research tools. The platforms are easy to develop on and while certain aspects (e.g. inter-application communication) are not currently possible, inter-portlet communication is not presently available in Gridsphere either (see, for example, Yang et al. 2006). SNSs can provide choice of tools and facilitate user-driven innovation in e-Research.

References

Ackland, R and R. Gibson. (2004): "Mapping Political Party Networks on the WWW," Proceedings of the Australian Electronic Governance Conference, 14-15 April 2004, University of Melbourne.

Ackland, R., O'Neil M., Bimber B., Gibson, R. and S. Ward. (2006): "New Methods for Studying Online Environmental-Activist Networks," paper presented to 26th International Sunbelt Social Network Conference, 24-30 April, Vancouver.

Barabási, A.-L., and R. Albert. (1999): "Emergence of Scaling in Random Networks," *Science*, 286, 509-512.

Barjak, F. and M. Thelwall. (forthcoming): "A statistical analysis of the web presences of European life sciences research teams", *Journal of the American Society for Information Science and Technology*.

González-Bailó, S. (2007): "Mapping Civil Society on the Web: Networks, Alliances, and Informational Landscapes," DPhil Thesis, University of Oxford.

Newman, M. E. J. (2002): "Assortative Mixing in Networks," *Phys. Rev. Lett.*, 89, 208701.

Shumate, M., and L. Dewitt (2008): "The North/South Divide in NGO Hyperlink Networks," *Journal of Computer-Mediated Communication*, 13, 405-428.

Thelwall, M., L. Vaughan, and L. Bjorneborn. (2005): "Webometrics," *Annual Review of Information Science and Technology*, 39: 81-135.

Wimmer, A. and K. Lewis. (2008): "Below and Beyond Racial Homophily. ERG models of a friendship network based on Facebook.com," Paper given at International Sunbelt Social Network Conference, St. Pete (Florida), January 22-27, 2008.

Yang, X, Wang, X and R. Allan. (2006): "JSR 168 and WSRP 1.0 - How Mature are Portal Standards?" Proc. WEBIST 2006 (WEBIST 2006), Setubal, Portugal, 11-13 Apr 2006.

8 Sampling is fundamental to empirical social science research but relatively foreign to those whose research involves searching over collections, since one would never take a sample of a collection (all members need to be locatable).