Back to JoSS
JoSS Article: Volume 12

A Relational Hyperlink Analysis of an Online Social Movement

Dean Lusher, dean.lusher@unimelb.edu.au
Psychological Science
University of Melbourne, Australia

Robert Ackland, robert.ackland@anu.edu.au
Australian Demographic and Social Research Institute
The Australian National University, Australia

Abstract: *In this paper we propose relational hyperlink analysis (RHA) as a distinct approach for empirical social science research into hyperlink networks on the World Wide Web. We demonstrate this approach, which employs the ideas and techniques of social network analysis (in particular, exponential random graph modeling), in a study of the hyperlinking behaviors of Australian asylum seeker advocacy groups. We show that compared with the commonly-used hyperlink counts regression approach, relational hyperlink analysis can lead to fundamentally different conclusions about the social processes underpinning hyperlinking behavior. In particular, in trying to understand why social ties are formed, counts regressions may over-estimate the role of actor attributes in the formation of hyperlinks when endogenous, purely structural network effects are not taken into account. Our analysis involves an innovative joint use of two software programs: VOSON, for the automated retrieval and processing of considerable quantities of hyperlink data, and LPNet, for the statistical modeling of social network data. Together, VOSON and LPNet enable new and unique research into social networks in the online world, and our paper highlights the importance of complementary research tools for social science research into the web.*

1. Introduction

Statistical analysis of hyperlink data has typically followed one of two broad approaches. First, techniques from network science have been used to identify structural properties such as power laws (Barabási and Albert 1999) in the degree distribution, where a small number of pages or sites receive the lions' share of inbound hyperlinks, with the majority receiving few or none. Hindman et al. (2003) argue that the existence of power laws on the web has implications for the visibility of different political messages, since search engines such as Google (which are important drivers of website traffic) generally rank better connected sites or pages more highly. Second, webmetrics (or webometrics) is an approach for analyzing hyperlink data that was originally developed for measuring scholarly or scientific activity using web data (see, for example, Almind and Ingwersen 1997; Björneborn and Ingwersen 2004; and Thelwall, Vaughan, and Björneborn 2005). A typical webmetric technique is ordinary least squares (or variants), where the counts of inbound hyperlinks are regressed on the characteristics of the websites and the actors who run the website (e.g. research team or organization), in an attempt to identify the attributes that lead to the acquisition of hyperlinks.

While social scientists are actively engaged in empirical analysis of hyperlink data, it is notable that the two common approaches for statistically analyzing hyperlink data originated in disciplines outside of social science: physics in the case of network science, and library and information science in the case of webmetrics. It is particularly curious that social network analysis (SNA), a sub-field of sociology that is focused on the representation and statistical analysis of social structures, has not been extensively used to analyze social structures on the web (represented via hyperlink networks).

However, the potential for using SNA to analyze hyperlink networks was first noted in the relatively early days of the web. Jackson (1997) considered that SNA "...has significant potential to generate insight into the communicative nature of Web structures" but argued that two of the core assumptions of SNA, the dependence of nodes within a network and the emergent property of networks, do not apply to the web. Further, the author was not comfortable with the nodes in a hyperlink network (pages or sites) being described as social actors and also appeared reticent to argue that the core SNA premise that structure of network relations impacts on the individual nodes and the system as a whole was applicable to the web, at least in formal terms. Thus, while Jackson (1997) felt that the structure of relations on the web would "...have important consequences for the way we communicate, and for what we understand as the structure of communication as a whole," the author was clearly less sanguine that formal SNA concepts and methods could carry over to the web. In contrast, Park (2003) had no reservations about describing websites as social actors and advocated that the analysis of hyperlink networks using SNA be called "hyperlink network analysis". Despite this early recognition of the potential of SNA for hyperlink analysis, there are not many examples in the literature where hyperlink data have been analyzed using formal SNA techniques. This is partly explained by the fact that there has not been much research providing theoretical justification for why a hyperlink network might be considered as comprising social actors, with behavior that influences (and is influenced by) other actors and the system as a whole, and thus suitable for analysis using SNA.[1] Thus, there is still a lack of clarity in the literature as to why SNA techniques might be used for studying hyperlinking behavior, and how such an approach might differ from the other two empirical approaches for studying hyperlink data.

In this paper we show that the analysis of a hyperlink network using SNA techniques is markedly different to the other approaches (in particular, webmetrics). We utilize a particular class of statistical models for social network analysis (SNA), named exponential random graph modeling or ERGM (Frank and Strauss 1986; Wasserman and Pattison 1996; Pattison and Wasserman 1999; and Robins, Pattison and Wasserman 1999) to explicitly test for the existence of "structural signatures" (Faust and Skorvetz 2002; Skorvetz and Faust 2002) in the hyperlink networks formed on the web for coordinated action.[2] We propose that the application of ERGM to hyperlink data be called "relational hyperlink analysis" (or RHA), and we contend that this approach is appropriate for modeling the behavior of actors who use hyperlinks in a relational manner, to fulfill particular social or organizational functions. Thus RHA is a relational social science framework, which pays particular attention to hyperlinks as social connections, not merely indicators of popularity or visibility.

We demonstrate RHA as a distinct approach for modeling hyperlink data using the example of an online social movement - the asylum seeker advocacy movement in Australia. The choice of an online social movement to illustrate RHA is based upon the expectation that these online actors exhibit the social and informal hyperlinking behavior that RHA is specifically designed to model. More specifically, as with Shumate and Dewitt (2008) and Ackland and O'Neil (forthcoming), we conceptualize advocacy groups as engaging in online collective behavior or mobilization. We contend that the main functions that are undertaken by Australian asylum seeker advocates are *research* into asylum seeker and refugee issues, *service provision* (e.g. provision of housing, health services, counseling and more to asylum seekers) and *lobbying* of the government or the UN. We hypothesize that asylum seeker advocates constructed a hyperlink network that was primarily designed to maximize the chances of favorable changes to legislation. In particular, we argue that the hyperlinking activities of refugee advocates were designed to raise the web presence or prominence of those groups specifically engaged in lobbying, so that web users could easily find these sites (either by following links or via search engines) to engage in direct political action (by signing petitions, attending rallies, etc.), and the submissions written by the lobby groups were easily found (via search). While we hypothesize this as the main goal of hyperlinking by advocacy groups – to maximize lobbying efforts – we also acknowledge that there will have been other reasons for web activity during this period.

Our final contribution relates to the use of advanced tools for empirical social network analysis using web data. Ackland (2009) has argued that empirical research using web data involves a wide range of specialized techniques and tools (encompassing web mining, text mining, data visualization, statistical social network modeling) and that it is not viable (or necessarily desirable) for these tools to be contained in a single piece of software. What is needed is a technological platform that will enable web researchers to easily access complementary software tools. E-Research (or cyberinfrastructure) technologies are designed to enable collaborative access to distributed research resources (data, methods, computational cycles) and hence should be able to provide such a platform. Our research into hyperlinking behavior on the web involves the use of two such complementary software programs, the VOSON System[3] (Ackland 2010), which is a tool for collecting and analyzing online networks, and LPNet (Wang, Robins, and Pattison 2006), used for the longitudinal statistical examination of social networks. While our joint use of these programs possibly does not formally constitute e-Research (VOSON and LPNet currently do not "talk to one another" via web services or grid technologies, which are hallmarks of e-Research), our research is a good example of how complementary tools can be used for advancing research into online networks and, hence, provides important insights for the development of cyberinfrastructure for social network research, and research more generally.

The structure of the paper is as follows. In Section 2, we introduce relational hyperlink analysis as a distinct approach for analyzing hyperlink data and also provide an introduction to ERGM. Section 3 provides background information on our empirical example – Australia's recent policies towards asylum seekers and refugees, and the activities of groups advocating on their behalf. Section 4 presents details on the data collection and preliminary analysis, and there is a comparison of VOSON with other related software. Section 5 presents a statistical analysis of asylum seeker advocacy hyperlink networks, using the LPNet software. In Section 6 we discuss the results of this analysis, and we present conclusions in Section 7.

2. Relational Hyperlink Analysis (RHA)

In this section, we first discuss the challenge of conceptualizing and identifying hyperlink networks as social networks. Given we have collected hyperlink network data that we can conceptualize as a social network, we then discuss how analysis should proceed. Using an example of a simple friendship network, we show that a given social network can be "unpacked" into various co-existing sub-structures and it is not straightforward to identify the social processes that may have led to the emergence of a given network. However, a relatively recent innovation in SNA, exponential random graph models (ERGM), is specifically designed to statistically unpack social networks, and we provide a brief introduction to this technique. Finally, we introduce *relational hyperlink analysis* (RHA) as the application of ERGM to hyperlink network data, and we compare RHA with webmetrics, a commonly-used approach for analyzing hyperlink data.

2.1 Social structures on the web

Social network analysis (SNA) is an approach for the analysis of social structures[4] that are formally represented as social networks (where nodes represent actors and ties represent the relationships between actors). A social network must be clearly defined if it is to provide an accurate representation of a social structure, and hence be useful for understanding how human social systems operate. As Laumann et al. (1983, p. 33, emphasis in original) suggest, “there is no sense in which social networks must ‘naturally’ correspond to social systems." The definition of a social network involves three fundamental and interrelated issues: (1) What constitutes a social tie? (2) Who are the nodes/actors? and (3) Where is the network boundary? These issues are not always explicitly thought through by the researcher, but as Laumann et al. (1983, p. 19) suggest, they should be given “conscious attention."

When we are studying networks on the web as representations of social structures, there is an even more pressing need for conscious attention to the tie-actor-boundary triumvirate. With regards to social ties, the Internet enables individuals and organizations to connect in many ways, for example via email, online chat groups and social network services such as Facebook.com. However, our interest here is in modeling hyperlinks between websites as social network ties. Suggesting that “a hyperlink is a hyperlink is a hyperlink” is as awkward as suggesting that “a tie is a tie is a tie." The general refutation of this mantra within the field of SNA indicates that social ties should be carefully defined either through the researcher’s refinement of a particular question, or by those within the context under study (e.g. sitting on a board, financial transactions, country borders). Considerable social network research suggests that tie type is important – for instance, the strong tie/weak tie argument (Granovetter 1973; Krackhardt 1992). Well-defined social networks may therefore distinguish, for instance, instrumental from expressive ties, positive from negative ties, or as noted, strong from weak ties. Different sorts of ties may function in different ways, and by combining all types of ties within a single network such subtleties may be missed and an understanding of how multiple networks intersect may not be taken into account. While it may be argued that linking to another site is a validation of that site, a link may represent a criticism or some other negative comment. In this sense, a hyperlink works in the exact opposite way in that it involves de-legitimizing another.

Similarly, reciprocal hyperlinks may represent disagreement rather than mutual legitimation. Further, the notion that “an enemy of an enemy is a friend” also suggests multiplexity of relation type, where a positive tie is dependent on the presence of two negative ties within a triad. While structurally identical to the notion that “a friend of a friend is a friend,” the meaning of these two triads is completely different. Not distinguishing types of ties can seriously change the interpretation of the structural pattern of social relations. Researchers need some way to define hyperlinks more acutely, and will not be able to answer more refined questions by assuming all hyperlinks are interchangeable. Therefore, fundamental distinctions (such as positive or negative relations) need to be taken into consideration when examining hyperlinks as social network ties. Finally, the issue of network sampling is perhaps a prime example of how the links between network actors actually define the network boundary, as actors are included in the network due to their ties with others. So, the selection of relations can have important implications for the network boundary specification.

The third concern is the problem of boundary specification, which we have noted necessarily entails the selection of nodes but also the type of social relation (Laumann et al., 1983). Is it acceptable to include as a node any website that may be connected to another in any way? "The realist strategy of setting network boundaries by definition assumes the proposition that a social entity exists as a collectively shared subjective awareness of all, or at least most, of the actors who are members" (Laumann et al. 1983, p. 21). So defining the actor set on the basis of a particular nodal attribute is the most common way of defining a boundary, as well as by participation in an activity or event is another (Laumann et al. 1983). In the case of asylum seeker advocacy groups, we are interested in those promoting change, and thus it is not enough to include websites with some content on asylum seekers (e.g. Department of Immigration, newspapers) who have direct control over policy or may have no particular view on the subject. However, the type of social tie may also have implications on boundary specification. For instance, supporter groups of a particular sporting team may be more likely to have positive social ties to one another than groups supporting those of competing teams. For asylum seeker advocates, a boundary may be drawn around any website involved in advocacy, though restricting it to groups in Australia tightens this specification by enforcing a geographical boundary. In short, when thinking about social network boundaries, the issues of nodes and relations must be taken into account, and it is clear that these questions provoke difficult considerations.

2.2 Unpacking social structures: An example of a simple friendship network

Consider the example of a friendship network in Figure 1, where node color refers to gender (yellow = male, blue = female) and size refers to age (larger nodes are older), and the arrows represent a directed friendship tie. In this network we see the presence of reciprocal ties and also transitive triads, which are common in friendship networks.[5] There are of course other network features here (see Figure A2 in the Annex for a more comprehensive but not exhaustive list). For the purposes of illustration, we presently focus on reciprocity and transitivity. Both reciprocity and transitivity are examples of *purely structural network effects*, which are defined as network effects involving ties that have nothing to do with actor attributes. In the case of a friendship network, reciprocity and transitivity occur because of social norms in friendship formation. In particular, one generally reciprocates when someone extends the hand of friendship, and the adage that "a friend of my friend is also my friend" is also a social norm. We do not assert that such patterns always happen, but the presence of such structures does not depend upon the characteristics of the individuals involved.

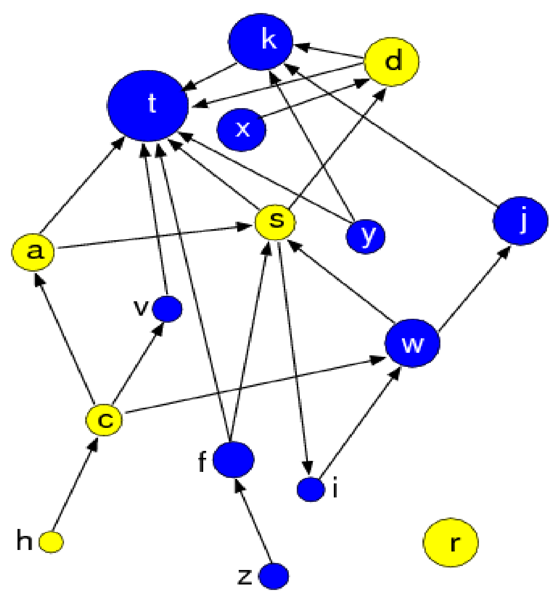


Figure 1: A social network of friendship relations

In contrast to purely structural network effects, there are *actor-relation effects*, which are network ties that are created because of the characteristics or attributes of actors.[6] Network effects (both purely structural and actor-relation) thus provide insight into the "structural processes necessary to explain how the network came to be" (Robins et al. 2009, p. 107). They tell us about consequential patterns of social relations, which in turn provide a window onto the social mechanisms which give rise to social relations (Hedström and Swedberg 1998).

In Figures 2a-2c we present three transitive triads that have been extracted from Figure 1. In Figure 2a, actor *y* nominates actor *k*, actor *k* nominates actor *t*, and the triad is closed by actor *y* nominating actor *t*. Similarly, in Figure 2b, *d* chooses *k*, *k* chooses *t*, and *d* chooses *t*. Further, in Figure 2c, actor *a* chooses *s*, *s* chooses *t*, and *a* chooses *t*. The problem we are faced with is determining why these particular triads have formed, and there are several competing explanations. For instance, the tie from actor *k* to actor *t* could be due to actor-relation effects, for example, actor *t* being older, or because actor *t* is also female (i.e. homophily). But *k*'s nomination of *t* could also be purely structural, with *k*'s decision being influenced by the fact that *y* nominates both *k* and *t* (this would be an example of *k* forming a transitive triad), or *t* being chosen because of a popularity effect (*k* deciding to nominate *t* because “everyone else does”).[7]

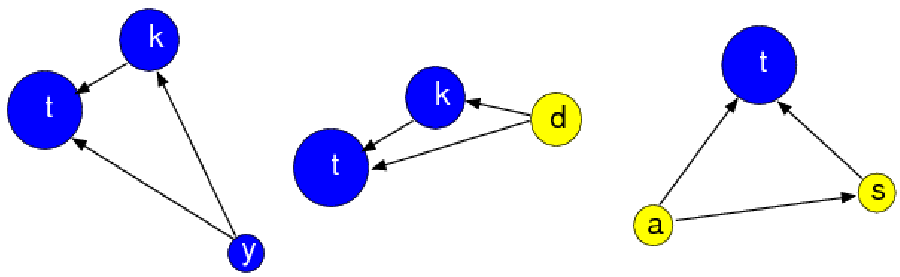


Figure 2: Three transitive triads in the friendship network

2.3 Exponential random graph models (ERGM)

Without information on the time sequence of tie formation, it is clearly very difficult to discern the reason why the above friendship network may have formed. With larger and more complex networks that are not easily visualized, the difficulty becomes even greater. More formally, any given observed network has a number of possible *realizations* ranging from a network in which no nodes are connected to that in which every node is connected to every other node. Monge and Contractor (2003, p. 49) note that “the statistical question of interest is why the observed realization occurred out of the rather large set of other possible graph realizations."

Statistical methods such as logistic regression can be used in an attempt to explain why a particular network has been realized (such an approach might be used to find the impact of node characteristics on the probability of a tie). However, this involves treating each tie as a unit of analysis and a standard logistic regression cannot be used since the assumption of independence of individual observations is violated (in the friendship network above node *c* links to 3 nodes (*a*, *v*, and *w*) and all of these ties will share the same error component). While robust standard errors can be used in such a situation (the point estimates are unbiased), the problem with standard logistic regressions is that there is no way of modeling the nature of interdependencies between ties (and as we saw with the friendship network above, there are theoretical reasons to expect particular types of interdependency).

One analytic approach of social network data that explicitly considers the interdependency of social ties is exponential random graph models (ERGM or *p** models).[8] ERGM are a particular class of statistical model for social networks that were originally proposed by Frank and Strauss (1986), and developed by Wasserman and Pattison (1996), Pattison and Wasserman (1999), Snijders, Pattison, Robins, and Handcock (2006), and Robins, Pattison and Wang (2009).[9] The ERGM class of models essentially works as a pattern recognition device, looking for consistencies in the ways social network ties are structured, as well as for associations between social network ties and individual attributes (Robins et al. 2001a; Robins et al. 2001b). These patterns are the network effects (or motifs) referred to above in the context of the simple friendship network: purely structural network effects and actor-relation effects. ERGM is fundamentally concerned with predicting network tie formation (Lusher, Koskinen, and Robins, in press), and so lends itself to also answering the question of whether social ties form due to the attributes of the nodes (*social selection*).

As outlined in Robins et al. (2007), all classes of ERGM have the general form in Equation 1.

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/k)\exp[\Sigma_A h_A z_A(\mathbf{x})] \qquad (1)$$

The components in Equation 1 are as such:

(i) **Pr(X=x)** is the probability of observing the graph, or network, that has been measured.

(ii) **(1/κ)** is a normalizing quantity which ensures that the equation is a proper probability distribution.

(iii) **exp** refers to exponential, hence *exponential* random graph models.

(iv) ***A*** represents a configuration, or network effect, included in the model, such as arc, reciprocity, or triad.

(v) Σ_A is the summation over all different configurations in the model.

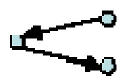
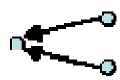
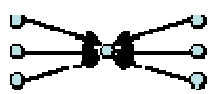
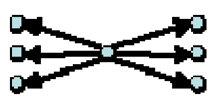
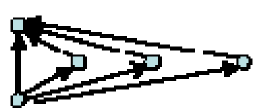
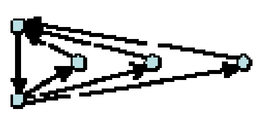
(vi) h_A is the parameter corresponding to configuration *A*.

(vii) $z_A(\mathbf{x}) = \prod_{x_{ij} \in A} x_{ij}$ is the *network statistic* corresponding to configuration *A* and is thus a count of the presence of configuration *A* in the observed network ($z_A(\mathbf{x}) = 1$ if the configuration is observed in the network **x**, and is 0 otherwise).

Equation (1) describes a general probability distribution of graphs and is used to determine the particular probability of observing a graph (or network). The specific probability of observing any graph [$\Pr(\boldsymbol{X}=\boldsymbol{x})$] depends upon both the network statistics [$z_A(\mathbf{x})$] and the non-zero parameters ($h_{\mathbf{A}}$) for all configurations *A* in the model. Configurations, or network effects, may include mutual ties, transitive triads, or more complex social structures. The presence of a configuration in a model does not imply that such a configuration is observed. Instead, configurations represent possibilities, and it is the network statistic $z_A(\mathbf{x})$ that tell us whether a particular configuration or structure is actually observed in a network. Of primary interest to many researchers are the parameter estimates ($h_{\mathbf{A}}$) which indicate the probability of the configurations from the observed network of interest. The model estimation produces parameter estimates and associated standard errors which, in a manner similar to standard regression techniques, are used to establish confidence in the estimation.[10] In essence, the parameter estimates of the configurations of the observed network are compared to those in a hypothesized distribution of networks of similar qualities, such as a similar number of nodes and a similar number of network ties. It is then possible to see if there are more or less configurations in the observed network than might be expected by chance. If there are some configurations occurring at greater or less than chance levels, it can be inferred that the observed network structures are not just coincidental observations, but consistent patterns of social relations. ERGM therefore allows the researcher to statistically identify various purely-structural and actor-relation network effects, and in the simple friendship example above, we mentioned a few of these possible network effects. The ability to control for purely structural self-organizing characteristics of social networks is an important advantage of ERGM. Not controlling for purely structural self-organizing network properties may lead to spurious actor-relation effects - that is, results may make it look like the qualities of actors are driving social tie formation when in fact it is purely structural self-organization.

Table 1 presents a more complete listing of purely structural network parameters which measure (and control for) endogenous or self-organizing structuring within the network.[11] The (1) arc parameter refers to the overall tendency of social actors to make social ties, while (2) reciprocity refers to the presence of mutual ties. The simple connectivity parameter (3) correlates the indegree and the outdegree, measuring the propensity of senders of ties to also receive them. Other effects account for (4) simple popularity and (5) more extensive popularity spread in the network, as well as (6) actor activity spread. There are also effects for (7) path closure (or transitivity), (8) cyclic closure, (9) multiple connectivity that does not result in closure, and (10) shared popularity of actors (for a more detailed description of these effects see Robins et al. [2009]).

Table 1: Purely structural network effects for ERGM

	Parameter	Image	Explanation	LPNet parameter name
1	Arc		One actor nominating another actor (baseline propensity to form ties)	Arc
2	Reciprocity		Mutual ties between two actors (models the tendency for reciprocation across the graph)	Reciprocity
3	Simple connectivity		Correlation of the indegree and outdegree, such that it models the propensity of senders of ties to also receive them	Mixed-2-star
4	Simple popularity		The propensity for a tie to be directed to an actor who is already active as a tie target (characterizing aspects of the indegree distribution)	2-in-star
5	Popularity spread		Indicative of the presence of highly nominated individuals within a network (models the indegree distribution)	K-in-star
6	Activity spread		Indicative of the activity of actors to engage many others (models the outdegree distribution)	K-out-star
7	Path closure		The propensity for ties to form as part of transitive triad or a multiply transitive configuration	AKT-T
8	Cyclic closure		The propensity for ties to form as part of a cyclic triad or a multiply cyclic configuration	AKT-C

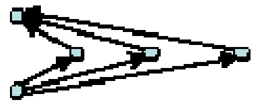
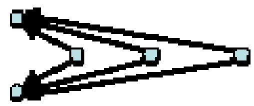
9	Multiple connectivity		The propensity for ties to form as part of formations involving multiple short paths between actors	A2P-T
10	Shared popularity		The propensity for popularity based structural equivalence involving multiple short paths between actors	A2P-D

Table 2 presents examples of actor-relation effects. S*ender* effects (1) reflect the impact of the presence (or absence) of a particular actor attribute on the propensity to send ties. A significant and positive sender effect indicates that actors with the attribute in question send more ties than expected by chance, while a significant and negative effect indicates that actors *without* the attribute send more ties.[12] R*eceiver* effects (2) work in a manner analogous to sender effects, except they reflect the impact of the presence (or absence) of a particular actor attribute on the propensity to *receive* ties. Lastly, the idea that *birds of a feather flock together* (McPherson, Smith-Lovin, and Cook 2001), otherwise referred to as assortative mixing, can be examined using the (3) homophily parameters, where a positive and significant parameter indicates that actors with a particular attribute are more likely than chance to send ties to other actors who share the same attribute.

Table 2: Actor-relation effects for ERGM

	Parameter	**Image**	**Explanation**	**LPNet parameter name**
1	Sender		The attribute of the sender of the tie, which may be continuous, categorical or binary (models the propensity of an actor with the attribute to send ties, i.e. to be active in network terms)	Rs

2	Receiver		The attribute of the receiver of the tie, which may be continuous, categorical or binary (models the propensity of an actor with the attribute to be popular)	Rr
3	Homophily		The propensity of a person with a binary attribute (e.g. “sex”) to choose other persons with the same attribute	Rb

Denotes actors with attribute.
Denotes actors with or without attribute.

A particular and important advantage of ERGM is the ability to specify particular *dependence assumptions* that accord with theory about how people form social ties in particular contexts. There are varying dependency assumptions, each with different degrees of complexity and realism. The simplest assumption, leading to what are termed Bernoulli random graph distributions, is where people form ties with others at a fixed probability α, thus independent of their other ties (Erdös and Renyi 1959). But such an assumption is not particularly realistic as, for example, in the case of sexual relations, at least some people are not likely to form a tie with another if they have an already existing sexual relation with another person. As such, there is likely to be some dependency in tie formation with respect to social relations. A more complex dependency assumption is dyadic independence which asserts that dyads, and not individuals, are independent. However, more complex dependencies were proposed by Frank and Strauss (1986), known as Markov dependence, which involve triads. Even more complex assumptions are made through realization (or social circuit) dependence (Pattison and Robins 2002; Snijders et al. 2006) which asserts the ways that four actors may be dependent upon one another. An example of realization dependence is the double-date. In the heterosexual case, two female friends interact with two male friends, and the relationship between one male and one female increases the possibility of interaction between the other male and female.

The selection of dependence assumptions leads to a particular specification of the model. Using the Hammersley-Clifford theorem[13] (Besag, 1974), it is possible to generate a probability distribution of random graphs using these configurations as the building blocks. This produces a range of networks of varying probability that are constructed from the pre-selected local social structures. “From a network perspective, individual behavior is viewed at least partially contingent on the nature of an actor’s social relationships to certain key others” (Laumann, Marsden, and Prensky 1983, p. 18). When we suggest that there are dependencies in the data, we do not mean between one variable and another (like age and eyesight) but within one variable (for instance, between the presence of one friendship relation and another).

2.4 Relational hyperlink analysis

We now introduce relational hyperlink analysis (RHA) as the use of ERGM to analyze hyperlink networks as social networks. It is important to note that RHA is not applicable for researching *any* hyperlink network. Rather, we propose that RHA is appropriate for studying the hyperlinking behaviors of social actors who *a priori* can be expected to exhibit both purely structural as well as actor-relation network effects. This point can be further clarified with a comparison of RHA with a commonly used collection of techniques for analyzing hyperlink data and website usage patterns, referred to as webmetrics. Webmetrics is an example of informetrics - a subfield of information science involving the use of mathematical-statistical approaches for the analysis of communication in science. A typical webmetric technique is ordinary least squares (or variants), where the counts of inbound hyperlinks to websites are regressed on the characteristics of the websites and the actors who run the website in an attempt to identify the attributes that lead to the acquisition of hyperlinks. In a recent example of webmetric research, Barjak and Thelwall (2008) regress 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 output indicators.

It should be emphasized that webmetrics comprises techniques other than counts regressions, but we focus on this technique since it is commonly used in this field and, further, it enables us to best distinguish RHA from webmetrics. However, it should be pointed out that counts regressions are also used in SNA, so we are not making a distinction here between webmetrics and SNA per se. Rather, our aim is to draw a distinction between a particular SNA technique (ERGM) and another statistical technique used both in SNA and webmetrics (counts regression), and show why the former is more appropriate for investigating certain types of behavior on the web.

In our above presentation of a simple friendship network, we distinguished two types of network effects: ties that occur for purely structural reasons (e.g. reciprocity and transitivity) and ties that occur because of the (exogenous) attributes of the nodes (e.g. homophily). A counts regression by definition ignores the fact that some ties may be purely structural and instead implicitly assumes that all ties are made for reasons relating to attributes of the actor receiving nominations. In contrast, ERGM acknowledges that ties might be made for purely structural reasons, as well as reasons relating to actor attributes, and provides a way of discerning the importance of each type of network effect. Following this, the simplest way of stating the difference between webmetrics and RHA is that, with webmetrics, the main question posed is "What are the qualities of actor receiving the most number of hyperlinks?" while RHA poses the more general question "*Why* do actors make or receive a hyperlink?"[14]

A counts regression approach is a more restricted approach than ERGM because purely structural network effects are omitted from the model. It is useful to understand why counts regressions are so central to webmetrics and why webmetricians have not investigated the use of the more general ERGM framework. We propose that webmetrics' implicit lack of recognition for the existence of purely structural drivers of hyperlink formation is due to the intellectual legacy of one of the main areas of informetrics, namely bibliometrics. Bibliometrics aims to quantitatively characterize and explain patterns of publication within academic fields. Webmetrics effectively treats hyperlinks as being analogous to an academic citation, and citation analysis typically does not allow for purely structural network effects, for both theoretical and practical reasons.

There are two broad theories that have been proposed to explain the determinants of citation flows (see, for example, Baldi 1998). One position is that citation is a normative process, where citations are used to recognize academic debt to authoritative and relevant prior work. In contrast, social constructivists disagree that academics follow internally sanctioned norms and instead argue that citations are mainly rhetorical tools of persuasion whereby authors attempt to buttress their arguments by making citations that are not based on academic merit or relevance, but because of the position or rank of the cited author in the field of research. Baldi (1998) tested these competing theories with a dataset of articles in an astrophysics research area, using a logistic regression where the probability of an article being cited was related to a content and quality of both the cited and citing article and the position or authority of the cited author in the stratification structure of science. The author found strong evidence that citations result from normative processes - the payment of intellectual debt - rather than social constructivist processes.[15]

The key point for the present paper is that neither of these competing theories of citation behavior involves purely structural network effects; both theories hold that citations are driven by characteristics of either the article or the author, and not by endogenous network effects. On a practical level, the unit of analysis in bibliometrics is either the article or the citation and the fact that an article can only cite another article that has already been published rules out, for instance, reciprocity as a potential driver of citations.[16] So while citation networks (where the nodes are articles and the ties are citations) can be regarded as socially constructed networks, they may not display some of the purely structural network effects that are present in social networks.

However, especially given the two main theories of citation behavior do not consider such purely structural network behavior as important, then counts regression approaches in bibliometrics appear to be justifiable.[17] We contend that this is why webmetrics - as an application of theories and methods from bibliometrics to the analysis of hyperlink data - does not involve empirical techniques that take account of purely structural network effects. Of course, it is also likely that ERGM, as a relatively new approach to relational data, has thus far slipped under the webmetrics radar.

The obvious next question is: Why this is important? The reason is that we expect a lot of hyperlinking activity does involve purely structural behavior, and standard webmetrics approaches (e.g. counts regressions) are not appropriate for studying the behavior of actors on the web in such circumstances. In particular, if there are purely structural hyperlinking behaviors that are not taken account of in the estimation approach, then the risk is that significance will be spuriously attributed to actor-relation effects. That is, we might mistakenly conclude that a particular attribute of the actors is important for network tie formation when, instead, it may simply be because there is an underlying purely structural network effect that has not been taken into account.

In conclusion, webmetrics is appropriate for studying particular types of hyperlinking behavior, for example the institutional or formal hyperlinking of government departments or where hyperlinks can be regarded as analogous to citations (e.g. research teams or universities). In contrast, we expect that social movement organizations will engage in more informal/grassroots networking behavior (i.e. social linking), and that there will be a certain amount of reciprocity and other purely structural network processes that must be controlled for in the analysis. We propose that RHA is appropriate for understanding the hyperlinking behavior of such social movement actors.

3. A Social Movement: Asylum Seeker Advocacy Groups in Australia

Information and communication technologies such as the web have had a major impact on the activities of advocacy groups. The web provides a low-cost way of espousing one's ideas, advertising, organizing events, mobilizing campaigns, sharing information, and engaging with like-minded others. It is a potentially rich information resource, an effective and economical means of communication, and appears to be a ready-made tool for political mobilization. While there is a large body of research into the use of the web for collective action and mobilization (e.g. Castells 1997; van de Donk, Loader, Nixon, and Rucht 2004), two recent studies are particularly relevant to the present paper. Shumate and Dewitt (2008) study 248 non-government organizations (NGOs) that are focused on HIV/AIDS, hypothesizing that the hyperlink network formed by these organizations is an example of an “information public good” that enables people to locate information and organizations working on this issue (by following links from other NGOs or else via search engines such as Google).[18] While Shumate and Dewitt (2008) use collective action theory (which in turn employs concepts from public choice theory), Ackland and O’Neil’s (forthcoming) analysis of the hyperlinking activities of environmental activists draws on the social movements literature, extending Diani’s (2003) network-conceptualization of a social movement to the online world. In particular, Ackland and O’Neil (2008) model actors in online social movements as engaging in online collective identity formation by using hyperlinks and website text to identify and highlight issues of concern.

One such online social movement is the asylum seeker advocacy movement in Australia. Australia’s policies towards refugees and asylum seekers have received much national and international attention (both positive and negative) over the past decade (European United Left/Nordic Green Left (GUE/NGL), 2005; Human Rights and Equal Opportunity Commission 2002, 2004; UNHCR 2004). It has been claimed by political commentators (Marr and Wilkinson 2003), by a prominent pollster (Roy Morgan Research 2005) and by two former prime ministers of Australia (Australian Broadcasting Corporation 2001) that the 2001 Australian federal election was won on the back of the government’s manipulation of asylum seeker issues. Specifically, the incumbent government’s use of fear, especially in implying that fraudulent refugees might arrive on Australia’s shores around the time of the September 11, 2001, World Trade Center bombings, was seen as instrumental in the incumbent’s political resurgence when the election seemed lost. However, the change in government in Australia in 2007 led to a dramatic shift in asylum seeker policies, with the policy redirection of 29 July 2008 realizing some of the hopes of many advocacy groups who had campaigned over the years for the better treatment of asylum seekers and refugees. While no doubt many factors were responsible for the new government’s outlook on asylum seekers, asylum seeker advocacy groups may have played a part in bringing the changes about.

While significant changes were made to Australia’s asylum seeker policies in early and mid 2008, before this time, under the purview of the Howard government, Australia’s policies were somewhat different. In 2006 Australia had a two-tiered refugee system that distinguished people fleeing persecution based upon their mode of arrival in Australia – a system that remains today though somewhat changed. On the one hand, Australia was (and remains) one of the few countries which have an annual quota for resettling refugees through the United Nations High Commission for Refugees (UNHCR) Program, indicating its proactive support of the UN Refugee Program (UNHCR 2004). Yet Australia was also regarded as having one of the harshest systems in the world for asylum seekers fleeing persecution who come directly to Australia’s shores (for a more detailed description of these policies, see

Lusher and Haslam 2007). Examples of the severe impact of Australia's policies include: the military intervention of the *Tampa*; the offshore processing of the 'Pacific Solution'; the sinking of vessel SIEV-X on its way to Australia and the loss of 353 lives in Australian waters, mostly women and children; the process of mandatory detention (i.e. detaining people indefinitely in prison-like conditions until their asylum claim is finalized, which has taken up to seven years in one case); and the mistaken incarceration and also deportation of mentally ill Australian citizens who were thought to be illegal immigrants. Further, rifts between the Australian and Indonesian governments in 2006 over the granting of asylum to 43 West Papuans resulted from what was seen as softening of government policy in response to considerable criticism from a government-implemented review of procedures (Palmer, 2005). One particularly notorious case was of five-year-old Shayan Badraie, who spent over twelve months in immigration detention, witnessing hunger strikes and suicide attempts. Shayan was diagnosed with acute and chronic post traumatic stress disorder that was attributed to his detention, which resulted in 70 trips to detention centre medical services, and eight visits to an external hospital. After detention, Shayan and his family were awarded refugee status.

While the terms asylum seeker and refugee are used synonymously in general discussion, they do in fact differ in meaning. An asylum seeker is a person who applies to the government of a country in order to be recognized as a refugee. By formal definition, a refugee is a person who "owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion, is outside the country of his nationality, and is unable to or, owing to such fear, is unwilling to avail himself of the protection of that country..." (*Convention relating to the Status of Refugees* 1951). Yet by highlighting this difference, Australian government policies portrayed asylum seekers coming directly to Australia as untrustworthy individuals who were not really suffering persecution and instead wished to take advantage of our generosity (Rodd 2007). This differential treatment on the distinction of these terms has received condemnation from the former UN Secretary General Kofi Annan (2004) who suggested that it broke the spirit of the Refugee Convention, and thus created a 'good refugee' and 'bad asylum seeker' distinction.

The response to Australian government policies on asylum seekers of this time was mixed. While many had criticized Australian government policies (Human Rights and Equal Opportunity Commission 2002, 2004), other countries such as Italy had begun to emulate the off-shore border protection system employed by Australia (European United Left/Nordic Green Left [GUE/NGL] 2005). Amongst the Australian public, there were a considerable number of Australians who supported government policy, as evidenced by the reelection of the Howard government to office in 2001. However, there was also a contingent of people who considered Australia's policies inhumane and against its international obligations as a signatory to the *Universal Declaration* *of Human Rights* (1948) and *Convention Relating to the Status of Refugees* (1951). In 2006, a proposed amendment to Australia's border control (Migration Amendment [Designated Unauthorised Arrivals] Bill 2006) aimed to scrap the Australian mainland as part of Australia's migration zone, so that all asylum seeking claims would have to be processed offshore (i.e. in another country). The online group GetUp! (getup.org) obtained 100,000 signatures against the proposed bill and tallied this in Parliament. Through lobbying of opposition and government ministers this proposed bill was scuttled by those advocating on behalf of asylum seekers in Australia. By 2008, there had been more sweeping changes to asylum seeker policy. While much of the harsh system remained, there was at least dialogue between the Australian government and asylum seeker advocatesaimed atfurther changes.

A fundamental question is whether there are patterns to the ways that asylum seeker advocacy groups hyperlink to one another that demonstrate coordinated political action, or is it random and lacking in coherence? Specifically, are groups that lobby for asylum seekers more linked to than other groups who support asylum seekers? To explore this, we examine the online social connectedness of Australian asylum seeker advocacy groups.

4. Data Collection and Preliminary Analysis using VOSON

This section describes the collection of the web data on asylum seeker and refugee advocates, and provides some preliminary descriptive analysis. The section begins with a brief introduction to VOSON, the tool that was used for the data collection and descriptive analysis.

4.1 VOSON: An e-Research tool for studying online networks

VOSON is server-based software that incorporates a web crawler [19], text analysis, network visualization and basic SNA techniques.[20] Users can access the software either via a web browser or via a plugin for NodeXL, which is an Excel 2007/2010 template for analyzing social media network data. [21] VOSON has been specifically designed for collecting and analyzing WWW hyperlink networks, that is, where the network nodes are web sites maintained by organizations or individuals, and the network ties are hyperlinks between these web sites. VOSON has been developed in the context of research in several areas including political party networks (Ackland and Gibson 2004), networks of political bloggers (Ackland 2005), the availability of information for migrants to Australia (Ackland and Gray 2005), and environmental social movements (Ackland et al. 2006; Ackland and O'Neil forthcoming).

There are other tools, aside from VOSON, that are used for the analysis of hyperlink networks. Of the tools that are publicly available and widely used, two deserve special mention (both of these tools have been around for longer than VOSON): Mike Thelwall's SocSciBot[22] (e.g. Thelwall 2009) and Richard Roger's Issuecrawler[23] (e.g. Rogers 2010). While we leave it to others to make a detailed comparison of the relative merits of VOSON, SocSciBot and Issuecrawler, there are two aspects that are relevant to the current paper. First, the tools have different disciplinary origins, and this will be reflected in the software features and how easily the tool can be used to solve a particular type of research problem. SocSciBot has been primarily designed as a tool for webmetrics, while the methods behind Issuecrawler appear to be derived from a more humanities-based perspective of the web (e.g. media and cultural studies). In contrast,

VOSON has been specifically designed for collecting inter-organizational hyperlink networks and analyzing these networks using SNA techniques.

The second feature that distinguishes VOSON is that it is an e-Research tool. E-Research (or cyberinfrastructure, as it is called in the U.S.) is the use of advanced information and communication technologies (ICTs) to enable new forms of collaborative research, involving access to distributed research resources (datasets, methods, compute cycles).[24] Based on this definition, SocSciBot is not an e-Research tool since it is client software that is downloaded on to the user's computer; there is no collaborative access of distributed research resources. Issuecrawler is a hosted service that is accessible via a web browser and thus clearly enables access to distributed research resources. However, it is not clear that there can be collaborative access to these resources (that is, is it possible for a team of researchers in different locations to access and work with a common dataset?). Unless collaborative access is allowed, then Issuecrawler is not an e-Research tool, as per the definition above.[25]

4.2 Refugee advocacy hyperlink network: Data collection using VOSON

An initial set of 67 seed pages was identified using Google searches and known asylum seeker advocacy group websites.[26] The seed pages are the entry pages to the sites of interest, e.g. the pages from which we expect to find links to other parts of the site, and where we expect there will be text explaining the main purpose of the site. The VOSON web crawler was then used to extract the outbound hyperlinks from the sites. Some of the seed websites were potentially very large and, for this reason, the crawler was set to crawl until: (1) 500 intrinsic (internal) pages were crawled; (2) 1,000 hyperlinks to other sites were found; or (3) 50 intrinsic pages had been crawled without the discovery of a new external hyperlink. The Google API was then used to find hyperlinks pointing to each of the seed pages, up to a maximum of 1,000 hyperlinks per seed page (this maximum is set by Google). The process of finding outbound and inbound hyperlinks resulted in a VOSON database containing 10,400 pages (including the 67 original seed pages). This initial data collection step was conducted in July 2006. Each of these 10,400 sites was manually examined by the researchers and included if they fulfilled the following criteria: (1) they advocated in some way on behalf of asylum seekers, and (2) they were located in Australia. This was a time-intensive process, but it was necessary since we needed to tightly define the network under study (as discussed in Section 2.1). We refer to all of these sites as advocates for asylum seekers and refugees, as we consider the presence of a website promoting asylum seeker and refugee issues an act of advocacy in itself. We conceptualize advocacy as incorporating direct and indirect action, petitions, and public education - and we see no need to differentiate these.

As noted for online networks, distinguishing types of ties from one another is difficult. Data mining strategies usually take any link from one site to another as evidence of a social relation. Without going to each link and coding its relevance we cannot distinguish between ties, as all types of relations are put together. There are informative issues that can be extracted from such analyses, where all types of ties are examined together, but the conflation of differing tie types may obscure the sorts of questions researchers are interested in. A possible way of getting around this dilemma is to manually examine every URL and classify it in a particular way. This is obviously extremely time-consuming and incommensurate with speedy data collection that data mining enables. Another potential way to restrict the range of types of ties is to carefully define a set of actors for the network. Implicit here is that the network boundaries and types of ties are interrelated. This involves the researchers manually checking each of the sites that are linked to and selecting only groups who adhere to the criteria set by the researchers. However, checking each site rather than each link is a much quicker process.

This process led to the identification of a final list of 211 seed pages. We note that some organizations use two or more hostnames (e.g. http://www.sievxmemorial.com/, http://www.sievxmemorial.org). In order to ensure that each organization's web presence was measured as accurately as possible, all known hostnames were included into the final seed list.

The VOSON crawler was then used to identify the outbound hyperlinks from the 211 seed sites, using the same web mining parameters described above. This time, only outbound hyperlinks were identified (inbound links were not collected using the Google API) because the analysis will be based on the hyperlink network formed by the seed sites. This second crawl was conducted in September 2006.

The second web crawl resulted in a database containing records for 21,861 pages: the 211 seed pages, plus the pages that these seeds linked to. The next data preparation step involved converting this database into a network dataset where each node represents the website of a refugee advocacy organization, and the ties represent hyperlinks between the websites. As mentioned above, several of the organizations have more than one hostname; the data preparation ensured that each organization was represented only once.

This data preparation step resulted in a network of 185 websites, however we excluded 41 of these to meet more tightly delineated inclusion criteria. Some were government departments or agencies involved in immigration matters such as the Department of Immigration and Citizenship (DIAC), which we excluded because it is not an advocate for change for asylum seeker policy but instead implements government policies regarding asylum seekers. Others were just advertising sites that had nothing to do with advocacy for asylum seekers. We also removed subsidiary state branches of international nodes (e.g. state branches of the Red Cross) because such sites will indubitably have hyperlinks to one another representing the formal connections of the organization, while we were studying the informal social linking behaviors of advocacy groups.

The final network dataset contained 144 websites and, because of the choice of these sites, we were relatively confident the ties expressed between these URLs were more likely to reflect some form of positive tie between the organizations running the websites. Further, it may be inferred that these would be instrumental ties, given we are talking about advocacy groups and social action. We still could not be sure how tight our definition of links was, given we are taking any tie between these groups. However, by excluding websites of the Australian government, of which many advocacy groups were critical, we removed some of the possibility for negative affect relations. This is a limitation of online data collection as we see it, and one to be overcome in the future, but one that we must live with at present and keep in mind in interpreting our results. Defining the network boundary also impacts our definition of a network tie. Given our focus on asylum seeker advocacy groups, we contend that hyperlinks to other like-minded sites are likely to reflect positive relations.

The final step of the data collection process involved re-crawling the 211 seed sites a final time in August 2008, forming a second database containing records for 36,314 pages. [27] Applying the same data processing steps as outlined above resulted in a second network dataset containing 144 websites.

4.3 Descriptive analysis

The two network datasets therefore provide information on the hyperlinking between the 144 seed sites at July 2006 (when the Howard government of Australia was in power) and August 2008 (a week after sweeping changes were made to asylum seeker policies by the relatively new Rudd government). Our use of longitudinal data allows us to make use of particular estimation routines which are better able to deal with data containing extreme degree distributions, and examine the dynamics of network tie formation. The attributes used in the analysis were from 2006 because LPNet is a model for the prediction of social ties, not for the prediction of actor attributes. No new sites were added at the 2008 time-point, and so the analysis is only on the presence of hyperlinks and their change over time.

We now present some of the descriptive analysis that is available via VOSON, focusing on the 144 seed sites in 2006. VOSON automatically classifies the seed sites on the basis of generic top-level domain (TLD) in the URL (e.g. .com, .edu)[28]; not surprisingly, the majority of the sites (85) are .org. The remaining sites are distributed as follows: .edu (24), .net (16), .gov (3), .info (2) and .asn (2). The generic TLD classification provides only limited information on the purpose or function of a particular site, so we examined each site in detail to determine key classifications of their goals and actions.

Our exploration identified three primary types of functions that asylum seeker advocacy groups engage in: lobbying, service provision, and research. *Lobby groups* lobby the media (via media releases) and also lobby the government via submissions to the government or to the UN directly. We contend that this active lobbying differs markedly from those who simply host a website calling for change. *Service providers* incorporate those groups who provide legal, health, education, counseling, food, accommodation, and/or employment to asylum seekers/refugees. Finally, *research* groups are those organizations that conduct research into asylum seeker and refugee issues. Websites were given a binary score on these three variables. It should be noted that these classifications are not mutually exclusive, so it is possible for an organization to lobby, provide services and conduct research. Some sites did not fall into any of these three major descriptors, and were seen as more general advocacy groups for asylum seekers in that their aims were to raise community awareness.

The classification of the 144 asylum seeker and refugee advocacy sites is presented in Table 3. The largest group is service (73 sites), followed by lobby (58 sites) and research (17 sites). Of note is that there are 29 sites here who are not involved in lobbying, service or research, but who are nonetheless advocates for asylum seekers and refugees. For the details of the websites and their attributes, see Annex Table A1.

Table 3: Characteristics of the 144 asylum seeker advocacy websites:

Cross-tabulations for Lobby, Service Provision and Research

Research		Service 0	Service 1	**Total**
0	lobby 0	27	41	**68**
	lobby 1	28	19	**47**
		55	**60**	**115**
1	lobby 0	12	5	**17**
	lobby 1	5	7	**12**
Total		**17**	**12**	**29**

The hyperlink network formed by the advocacy groups in 2006 has a density (the number of hyperlinks as a proportion of the possible number of hyperlinks) of 0.046. The average seed site made 6.6 hyperlinks to other seeds; lobby sites received more hyperlinks than average (around 8.9 per site) and this constitutes preliminary evidence that lobby groups are more prominent within this network, in the sense that other actors appear to be actively directing people to these sites (via hyperlinks). In the next section this is further investigated using statistical methods. Further information on the degree distributions is presented in Figures 3 and 4, which show the 2006 hyperlink network of asylum seeker advocacy groups, where the nodes are arranged along the vertical axis in order of increasing indegree (Figure 3) and outdegree (Figure 4). The lobby groups are the red nodes and it is notable that of the four top-ranked nodes in Figure 3, three of these are lobby groups.

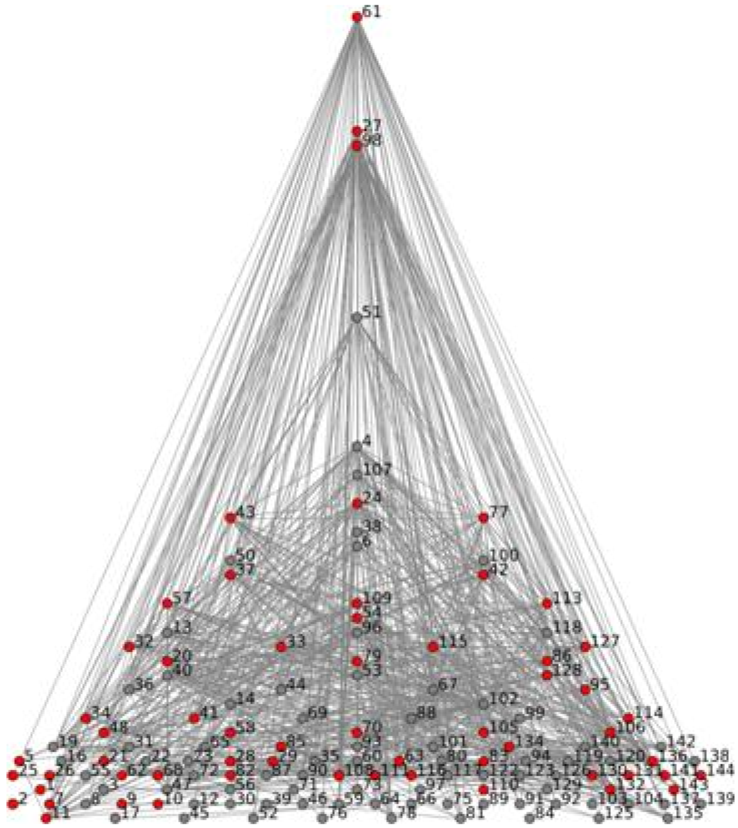


Figure 3: Hyperlinks between asylum seeker advocacy groups, sorted hierarchically by indegree nominations, 2006 (red nodes are lobby groups)

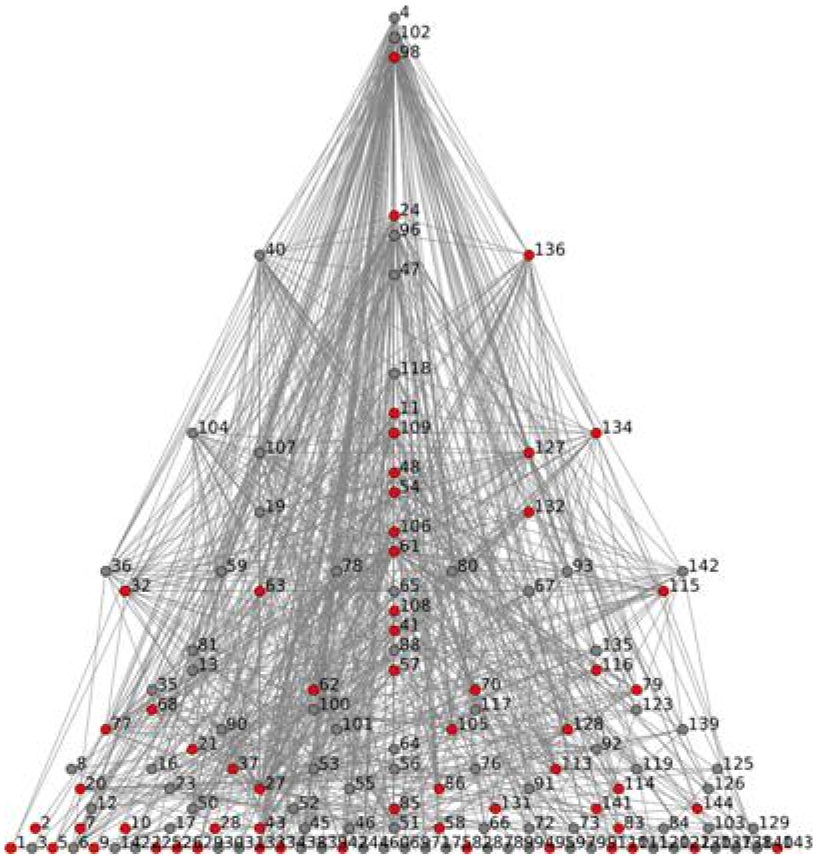


Figure 4: Hyperlinks between asylum seeker advocacy groups sorted hierarchically by outdegree nominations, 2006 (red nodes are lobby groups)

While the hierarchical maps in Figures 3 and 4 are useful for identifying nodes with large indegree/outdegree nominations, they do not reveal community structure or clustering of the sites. There are many ways for visualizing clustering in networks; Figure 5 shows the asylum seeker advocacy hyperlink network, drawn using the LinLogLayout force-directed graphing (FDG) layout of Noack (2007), where the lobby groups are indicated by red nodes and node size is proportional to indegree.[29] A screenshot of VOSON with the FDG and a cross-tabulation is shown in Figure A4 in the Annex. One thing to note from Figure 5 is that the lobby sites are fairly evenly distributed throughout the network, indicating that they are receiving hyperlinks from (and making hyperlinks to) the other two types of actors that we have identified. There is a small cluster of sites in the bottom right-hand corner of the map (indicated by the green dotted line) that are primarily service organizations.

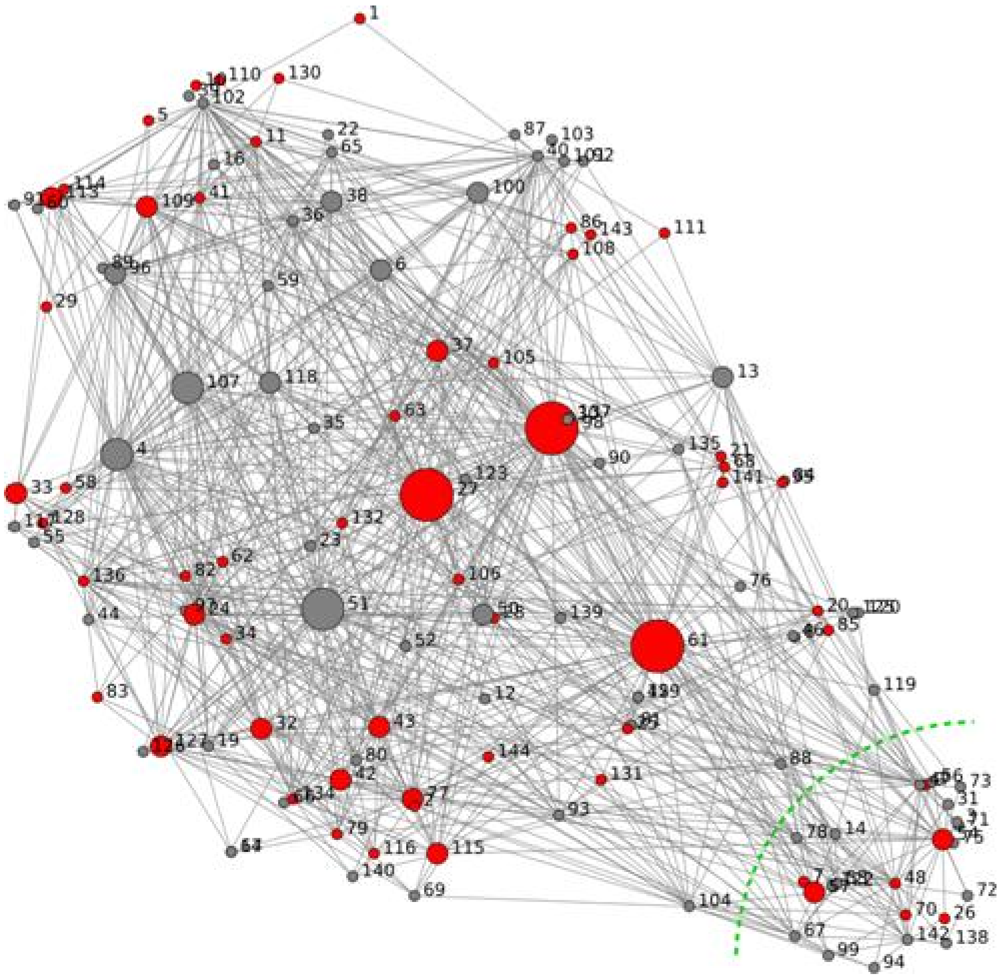


Figure 5: Force-directed map of hyperlink network for 136 (non-isolate) seed sites, 2006
(red nodes are lobby groups, node size proportional to indegree)

VOSON collects page meta keyword data (keywords describing the main focus or purpose of a website, often embedded into the HTML so as to ensure appropriate ranking by search engines) and text content extracted from the body of the web page. While the web crawler extracts hyperlinks by crawling (where possible) the entire site, text data was only extracted from the seed pages. As noted by Ackland and O'Neil (forthcoming), collecting text data only from the top-level page reflects both pragmatism regarding data storage capacity (some of the sites contain thousands of pages) and a view that an organization will place statements that best describe its activities or mission on the homepage, rather than buried deep within the site. While the text data were collected from the asylum seeker and refugee advocacy seed sites, we do not present text analysis in this paper (see Ackland and O'Neil forthcoming, for an example of website text analysis in the context of online social movements).

5. Relational Hyperlink Analysis using LPNet

There are three main software packages for conducting ERGM: a suite of tools collectively referred to as PNet (Wang et al. 2006), StocNet (Snijders et al. 2008) and statnet (also known as ergm: Hunter et al. 2008). Each of these software packages has its particular strengths, but we utilized the PNet suite due to its familiarity to us (one of the authors works within the research team in which it was created).

Our initial attempt to estimate an ERGM for the 2006 refugee advocacy hyperlink network only involved the use of the PNet tool, which was the first tool developed in the PNet suite, and is designed for the simulation and estimation of social selection ERGM for network data collected at a single time-point.[30] However, we were not able to produce a convergent model (i.e. produce stable parameter estimates) for the 2006 data. It is well-known that the presence of high degree nodes can cause convergence problems for ERGM, and Figures 3 and 4 clearly show the existence of such outlier nodes with very large indegree and outdegree nominations. The presence of outlier nodes presents difficulties for obtaining maximum likelihood estimates that generate a graph distribution centered upon the observed network (the graph space is extraordinarily large and the sampling thereby involves an enormous number of graphs). The application of exogeneity constraints to the model (i.e. fixing the ties for the outlier nodes, and modeling the rest of the network) still did not result in a convergent model.

To address the problem of non-convergence, we used the 2008 hyperlink network data and modeled both networks with LPNet (longitudinal PNet). With network data at two time-points, model convergence is easier to achieve, as the focus on changes to the first observation rather than explaining each tie variable unconditionally reduces the network space we need to explore. With LPNet, the first network is held constant, but it is not simply a matter of modeling the second network given the first (such as modeling one network in relation to another exogenous covariate network). Instead, while the first network is held constant we model the process of getting from the first to the second (more recent) network. As such, even if there was, for example, a high level of reciprocation in both of the networks presented here, if there was no change in the number of reciprocated ties, then our models would not produce a significant effect for reciprocation. In contrast, with cross-sectional ERGM we cannot model change because we have only one network. ERGM works as a static pattern-recognition device, determining if in the network we have observed, we see more of certain network structures than we might expect to see by chance. ERGMs for cross-sectional networks examine static network structures, not changes in network ties (though the underlying assumption is that those static structures are traces of dynamic processes).

The model assumed in LPNet is tie-based in the sense that, conditional on the event that a tie-variable is updated, we model the conditional probability of a tie being present, conditional on everything else, in a "ERGM-like" way

(Snijders and Koskinen, forthcoming). As described in Snijders and Koskinen (forthcoming) and Snidjers (2006), a tie-based model may be phrased in terms of an actor-oriented model (Snijders 2001; Snijders, Koskinen, and Schweinberger 2010) that may be filled using SIENA; but here the assumptions of the actor-oriented model are relaxed. Whereas the change in a tie in an actor-oriented model is modeled as a result of action taken by one or both of the actors involved in the dyad, we do not make that assumption explicitly here.

Additionally, as LPNet assumes a social selection model, any attributes that you use are for the first network. So in LPNet actor attributes are fixed, and it is the ties that are predicted. SIENA is able to simultaneously model both social influence and social selection processes (or the coevolution of network ties and actor attributes; see Steglich, Snidjers and Pearson, forthcoming).

5.1 Results

The first step of building the model using LPNet was the selection of the purely structural network effects to be included as controls. The inclusion of purely structural network effects caters for interdependency among the observations and enables valid inference about actor-relation network effects (our primary focus). Model convergence and goodness of fit (GOF) statistics are used to guide the choice of structural network effects, but the experience of the researcher in analyzing similar networks is also important.

Table 4 shows the purely structural effects that we included in the model. We decided to include two separate multiple connectivity parameters (A2P-T and A2P-D) rather than a single joint parameter (A2P-TD) because the valences of these two effects are different. Further, the use of two popularity parameters, the 2-in-star (a Markov parameter) with the K-in-star (a higher order parameter: alternating *k-in-stars*), is useful when the indegree distribution is highly skewed, as is the case with the asylum seeker advocates network, where there are some very popular network actors. In less complex and skewed social networks, the K-in-star parameter on its own might be sufficient. In a very simple social network the 2-in-star might be sufficient instead of its higher order counterpart, the K-in-star. The window in the LPNet GUI demonstrating the selection of purely structural parameters is shown in Figure A2 in the Annex.

As shown in Table 4, we used three actor-relation parameters (sender, receiver and homophily) for each of the three actor-relation effects of interest (lobby, service and research), resulting in nine separate actor-relation network parameters (see Figure A3 in Annex for screenshot of LPNet).

Table 4: Longitudinal ERGM parameter estimates (and standard errors) for Lobby, Service and Research groups (for 144 nodes at two time-points: 2006 and 2008)

Parameter	Estimate (SE) Model A	Model B
Purely structural effects		
Arc	-3.16 (0.19) *	-6.43(0.61) *
Reciprocity		1.46 (0.32) *
Simple Popularity (2-in-star)		0.09 (0.01) *
Popularity spread (K-in-star)		-0.04 (0.26)
Activity spread		1.06 (0.23) *
Path closure (AKT-T)		1.13 (0.11) *
Cyclic closure (AKT-C)		-0.23 (0.08) *
Simple connectivity		-0.01 (0.01)
Multiple connectivity (A2P-T)		-0.03 (0.03) *
Shared popularity (A2P-D)		0.033 (0.015) *
Actor-relation effects		
Homophily effects		
Lobby	-0.12 (0.28)	-0.16 (0.35)
Service	0.01 (0.32)	-0.07 (0.29)
Research	0.82 (0.39)*	1.08 (0.41) *
Sender effects		
Lobby	0.23 (0.20)	0.26 (0.21)
Service	-0.74 (0.26)*	-0.75 (0.24) *
Research	0.62 (0.18) *	0.32 (0.18)
Receiver effects		
Lobby	0.39 (0.19)*	-0.02 (0.20)
Service	0.07 (0.15)	0.04 (0.17)
Research	-0.34 (0.27)	-0.43 (0.27)

In Model A, we run a Bernoulli model in which the only structural parameter is the Arc, but which still includes the three actor-relation effects. As noted previously, such a model assumes that the presence of one social tie is independent of the presence of another. This model gives us a general sense of how network ties are being made with regard to the actor attributes of interest, but is incomplete because it does not account for purely structural effects. However, it is useful because it provides comparative results to Model B, which includes both purely structural and actor-relation effects.

A parameter estimate greater than (in absolute value) two times the standard error is regarded as demonstrating a major effect. A significant and positive effect for a parameter indicates that it occurs at greater than chance levels, *given the other parameters in the model*. A significant and negative parameter estimate indicates that it occurs at less than chance levels, *given the other parameters in the model*. We stress the 'given the other parameters' to indicate the interdependency of the parameters in ERG models. For instance, a model exploring friendship ties that includes two parameters, arc and reciprocity, may find a significant and negative effect for arc and a positive and significant effect for reciprocity. The results need to be interpreted together, and indicate that there few nominations of others in the network (negative arc effect) outside of, or unless, they are reciprocated (positive reciprocity effect).

All parameters in our model indicated adequate convergence of the Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE) algorithm. To examine how well our model fits the data, we have used the goodness of fit (GOF) within the LPNet program.[31] The model did not fit for Model A but there was a good fit for Model B, with all parameters included in the model less than 0.1, and all other non-included parameters less than 2.0, including the in- and outdegree distributions.[32] By contrast, there are many parameters in the GOF for Model A where non-observed parameters (e.g. reciprocity) are >2.0 (in fact reciprocity is 6.134 in Model A) and therefore are not well captured by this model A.

For the actor-relation effects in Model A, we see a significant and positive homophily effect for the research groups, indicating that they are likely to link to other research groups. There is a negative and significant sender effect for service, indicating that service provider groups are less likely to make hyperlinks to other websites than might be expected by chance. However, the positive and significant sender effect for research indicates that they make many links to other websites. Finally, there is a positive and significant receiver effect for lobby. This indicates that lobby groups receive more ties than expected by chance, given all other parameters in the model. The conclusion then from Model A is that there is an overall tendency for groups lobbying for asylum seekers to receive many hyperlinks. Model A thus provides support for our hypothesis that lobby groups are the most prominent within the overall asylum seeker sector.

However, the inclusion of purely structural parameters (Model B) leads to a different conclusion. We reiterate that Model A does not incorporate complex dependency assumptions between network actors and is primarily concerned with the effect of actor attributes on social tie formation. In contrast, Model B examines exactly the same actor-relation effects as Model A, but also takes into consideration complex interdependencies in the data and the ways in which social ties arise for purely structural self-organizing reasons. In examining Model B, most importantly, there is no longer a significant and positive receiver effect for lobby groups. There is still a homophily effect for research groups, indicating that research groups have a greater propensity to hyperlink to other research groups. There is only one significant sender effect – negative for service groups, indicating they do not send many hyperlinks. Finally, there are no significant receiver effects for any of the attributes.

The purely structural parameters also add some interesting elements to the story in their own right. We find that there are significant and positive effects for reciprocity, path closure (AKT-T, transitive clustering), popularity (2-in-star, simple) and activity spread. There are significant and negative effects for cyclic clustering and transitive multiple connectivity, meaning that we see less of these particular network formations within this network than expected by chance. Also, there is a positive and significant effect for shared popularity, indicating that two sites are selected at greater than chance levels by many other sites, but do not link with one another. The simple connectivity parameter is not significant. Simple connectivity is a measure of the correlation of the indegree and outdegree, and so this result indicates that those sites that send many ties are not those that also receive many ties, given the other effects in the model. The purely structural effects taken together are suggestive of considerable hierarchy in the structure of hyperlinks. The transitive clustering and shared popularity parameters demonstrate very hierarchical structures, as does the popularity spread effect. The significant activity spread effect suggests that hyperlinks are not costly as there are a number of sites making many links to other sites.

6. Discussion

LPNet was used to statistically analyze the Australian asylum seeker advocacy hyperlink network, using 2006 and 2008 data collected by VOSON. Model A explored the propensity of websites to send and receive ties primarily based on actor-level attributes, whereas Model B also controlled for purely structural self-organizing network configurations which are known to be present in human social networks. Importantly, the results demonstrate that an assumption of independently forming ties for this advocacy hyperlink network (Model A) is not tenable, and that we need to account for more complex dependencies in social ties through higher-order purely structural effects (Model B). The inclusion of several purely structural variables (most of which are significant) ensures we do not overestimate the importance of the sender and receiver effects, and led to the disappearance of the receiver effect for the Lobby group that had been found in Model A.

The inclusion of the purely structural variables thus leads to a fundamentally different understanding of the advocacy hyperlink network than was gained via the descriptive statistics presented in Section 4 and the ERGM results in Model A. In particular, while we found that Lobby groups receive a higher-than-average number of indegree nominations and that they also have a significant receiver effect after controlling for all actor-relation effects (but not purely structural effects), Model B indicates that an apparent propensity for Lobby groups to receive many ties is in fact explained by purely structural effects (such as reciprocity, path closure and popularity effects). This suggests that counts regressions using hyperlink data, which is akin (but not exactly the same) to what was done in Model A, can produce potentially misleading results. In short, there is a need to control for the dependencies in the social ties via the inclusion of higher order purely structural network characteristics.

From the visualizations and an examination of the indegree counts it is clear that some lobby groups are extremely popular within this online advocacy network. The results of Model B do not contradict the finding that lobby groups are the most central or popular websites. The difference between Model A and Model B is what accounts for this popularity – that is, what is the *social process* that leads to such ties? The results of Model B allow us to argue that the prestigious sites in this hyperlink network are there because of purely structural tendencies in social tie formation. That is, these sites are popular because they are located in parts of the network where there is high clustering, or high reciprocity. It is the purely structural aspects of the network that explain popularity, not the particular attributes of the sites.

Thus, our answer to the question of whether asylum seeker advocacy groups are organized in a way to direct site visitors (via hyperlinks) to those lobbying on behalf of asylum seekers varies dramatically when we include higher order parameters in our longitudinal exponential random graph model to control for purely structural explanations for social tie formation. To help explain this we can use an analogy of understanding why someone is a billionaire. Counting ties just tells us if someone is a billionaire or not, but says nothing of the processes that led to the person becoming a billionaire, for example, whether it was by inheritance or individual ability, or by both. Incorporating actor relation and purely structural effects into the model does not change someone’s billionaire status, but may enable us to better understand how it came about.

So, while people *are* being directed to lobby groups via the hyperlink network formed by the asylum seeker advocacy sites (lobby sites are the highest indegree nodes in the network), we are not able to detect a concerted effort by the advocacy sites to hyperlink to the lobby sites. Some sites that just happen to be run by lobby groups have many ties because of social norms in social relations such as reciprocity or transitivity (a friend of a friend is a friend). Or there may be other non-measured characteristics that explain why hyperlinks are present. But importantly, it is not because they are lobby groups that they are popular.

The use of ERGM revealed that the hyperlink network exhibits a number of characteristics of a *social* network; in particular, reciprocity, transitivity, and homophily found in many human social networks. The fact that the asylum seeker advocacy hyperlink network does appear similar to other human social networks justifies our use of RHA, as opposed to webmetrics. It would be very useful to further explore the connection between the online and offline worlds in relation to social connections between these groups.

The asylum seeker advocacy hyperlink network does, however, differ from offline social networks in two ways: network expansiveness and popularity, demonstrating considerable star-like nominations in the network. This suggests that network nodes are not economical in how they form social ties to others, leading to some websites making a large number of links to other sites, and some websites receiving many links from others. This indicates support for the general conception that online social ties may be (relatively) cost free. Another important purely structural difference is the significant and positive shared popularity effect for (A2P-D). This parameter represents the propensity of a number of websites to link to two specific websites, but also that there is not necessarily a hyperlink connection between these two popularly selected sites. This particular structural effect is not generally seen in social networks - there are usually links between these popular nodes, resulting in transitivity. This may not be the result of our use of hyperlink data, but may reflect something in the "real world" relationships between these organizations. In offline settings, such an effect is often interpreted as suggesting some form of factionalization or friction within the network.

Finally, we found that longitudinal ERGM is better able to deal with some of the difficulties of online data, namely extreme degree distributions. A convergent model was achieved relatively easily when we used LPNet with two data points. This does suggest that longitudinal modeling can overcome extreme degree distributions, something it seems which may be a common characteristic of hyperlinked social networks.

7. Conclusion

In this paper we identified relational hyperlink analysis as a distinct approach for empirical research into hyperlink networks, and compared this approach with webmetrics. We contend that RHA is appropriate when there is an expectation that actors are using hyperlinks in an informal manner, that is, where the hyperlink network is expected to exhibit characteristics that are often found in social networks. Our study of the hyperlinking behavior of Australian asylum advocacy groups provided strong justification for the use of RHA. We found that the hyperlink network *does* exhibit many of the characteristics of a social network. Further, we would have made incorrect conclusions regarding the underlying reasons for hyperlinking behavior of the advocacy groups (in particular, their tendency to hyperlink to lobby groups) if we had used a counts regression (a common webmetric approach), rather than RHA.

It is important to note that our paper should not be regarded as an attack on webmetrics. Rather, our main message is that webmetrics may be a useful approach for studying particular phenomena on the web, for example the formal institutional linking of government agencies, but is not well suited for analysis of more informal, social behavior where websites may be seen as representing social actors (e.g. social movement organizations). We propose that RHA is appropriate for research into the types of actors on the web for whom hyperlinks have “intrinsic value and serve to promote some ideas, people, and organizations over others” (Shumate & Dewitt, 2008, p. 407, in reference to the “The Hyperlinked Society” conference of 2006).

Our paper also highlighted the importance of research tools for social science research into the web. The VOSON software provides a means of retrieving and preparing considerable quantities of hyperlink data that, if done manually, would be extremely time-consuming. In pairing this data collection tool with the software for statistical models for social networks, namely LPNet, a powerful combination of tools arises. Together, VOSON and LPNet enable research into social networks in the online world in unique ways. There are a number of possibilities for this combination of tools to understand how the web is structured and utilized, and what we can learn about issues online. For instance, with LPNet we clearly see that when purely structural network effects are not taken into account that our interpretations of social tie formation across the network may be inaccurate and lead to incorrect conclusions about the social processes underlying hyperlinks. In this particular substantive case we would have concluded that asylum seeker advocacy groups were informally coordinated in directing people to websites lobbying for change when with more principled investigations of the network data there is in fact no evidence for such an explanation.

References

Ackland, R. (2005). "Mapping the U.S. Political Blogosphere: Are Conservative Bloggers More Prominent?" Refereed paper presented to BlogTalk Downunder 2005, 19-22 May, Sydney.

Ackland, R. (2009). "Social Network Services as Data Sources and Platforms for e-Researching Social Networks." Special Issue on e-Social Science: *Social Science Computer Review* 27, 4 (Winter 2009): 481-492.

Ackland, R. (2010). "WWW Hyperlink Networks." In D. Hansen, B. Shneiderman and M. Smith (eds.), *Analyzing Social Media Networks with NodeXL: Insights from a connected world* (San Francisco: Morgan-Kaufmann).

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

Ackland, R. and E. Gray (2005). "What Can Potential Migrants Find Out About Australia from the WWW?" *People and Place* 13, 4: 12-22.

Ackland, R. and M. O'Neil (forthcoming). "Online Collective Identity: The Case of the Environmental Movement." *Social Networks*.

Ackland, R., M. O'Neil, B. Bimber, R. Gibson 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.

Almind, T. and P. Ingwersen (1997). "Informetric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics.'" *Journal of Documentation* 55, 5: 404-426.

Annan, K. (2004). "January 29, 2004: United Nations Secretary-General Kofi Annan's Address to the European Parliament upon receipt of the Andrei Sakharov Prize for Freedom of Thought (Brussels)." Available: http://www.europa-eu-un.org/articles/en/article_3178_en.htm. [May 2011]

Australian Broadcasting Corporation (2001). "Broadcast 14/11/2001: Fraser Blasts Asylum Seeker Policy." Available: http://www.abc.net.au/lateline/content/2001/s417232.htm. [March 2011]

Baldi, S. (1998). "Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model." *American Sociological Review* 63, 6: 829-846.

Barabási, A.-L. and R. Albert (1999). "Emergence of Scaling in Random Networks." *Science* 286, no. 5439: 509-512.

Barjak, F. and M. Thelwall (2008). "A Statistical Analysis of the Web Presences of European Life Sciences Research Teams." *Journal of the American Society for Information Science and Technology* 59, 4: 628-643.

Besag, J. (1974). "Spatial Interaction and Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society Series B-Methodological* 36, 2: 192-236.

Björneborn, L. and P. Ingwersen (2004). "Toward a Basic Framework for Webometrics." *Journal of the American Society for Information Science and Technology* 55, 14: 1216-1227.

Castells, M. (1997). *The Power of Identity* in volume 2 of the series *The Information Age: Economy, Society and Culture* (Oxford: Blackwell).

Contractor, N. S., S. Wasserman and K. Faust (2006). "Testing Multi-theoretical Multilevel Hypotheses about Organizational Networks: An Analytic Framework and Empirical Example." *The Academy of Management Review* 31, 3: 681-703.

Convention Relating to the Status of Refugees (1951). Available: http://www2.ohchr.org/english/law/refugees.htm. [May 2011]

Diani, M. (2003). *Networks and Social Movements: A Research Programme.* In M. Diani and D. McAdam (eds), *Social Movements and Networks: Relational Approaches to Collective Action* (Oxford: Oxford University Press).

Erdös, P. and A. Renyi (1959). "On Random Graphs." *Publicationes Mathematicae Debrecen* 6: 290-297.

European United Left/Nordic Green Left (GUE/NGL) (2005). *Lampedusa and Mellila: Southern Frontier of Fortress Europe*. Available: http://www.guengl.eu/upload/EN_Lampedusa.pdf. [May 2011]

Faust, K. and J. Skvoretz (2002). "Comparing Networks across Space and Time, Size and Species." *Sociological Methodology* 32, 1: 267-299.

Frank, O. and D. Strauss (1986). "Markov Graphs." *Journal of the American Statistical Association* 81, no. 395: 832-842.

Fulk, J., A. Flanagin, M. Kalman, P. Monge and T. Ryan (1996). "Connective and Communal Public Goods in Interactive Communication Systems." *Communication Theory* 6, 1: 60-87.

Granovetter, M. S. (1973). "The Strength of Weak Ties." *American Journal of Sociology* 78, 6: 1360-1380.

Hedström, P. and R. Swedberg (1998). *Social Mechanisms: An Analytical Approach to Social Theory* (New York: Cambridge University Press).

Hindman, M., K. Tsioutsiouliklis, and J. A. Johnson (2003). "Googlearchy: How a Few Heavily Linked Sites Dominate Politics Online." Paper presented at the annual meeting of the Midwest Political Science Association. Available: http://www.matthewhindman.com/images/docs/mpsa03.pdf. [May 2011]

Human Rights and Equal Opportunity Commission (2002). "Transcript of Hearing [of National Inquiry into Children in Immigration Detention] - Sydney - Tuesday 16 July 2002." Available: http://www.hreoc.gov.au/human_rights/children_detention/transcript/sydney_16july.html. [May 2011]

Human Rights and Equal Opportunity Commission (2004). "A Last Resort? National Inquiry into Children in Immigration Detention." Available: http://www.humanrights.gov.au/human_rights/children_detention_report/report/index.htm. [May 2011]

Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris (2008). "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24, 3. Available: http://www.jstatsoft.org/v24/i03/. [May 2011]

Jackson, M.H. (1997). "Assessing the Structure of Communication on the World Wide Web." *Journal of Computer-Mediated Communication* 3, 1. Available: http://jcmc.indiana.edu/vol3/issue1/jackson.html. [May 2011]

Krackhardt, D. (1992). "The Strength of Strong Ties: The Importance of Philos in Organizations." In N. Nohria and R. G. Eccles (eds.), *Networks and Organizations: Structure, Form and Action* (Boston: Harvard University Press), 216-239.

Laumann, E. O., P. V. Marsden, and D. Prensky (1983). "The Boundary Specification Problem in Network Analysis." In R. S. Burt and M. J. Minor (eds.), *Applied Network Analysis* (London: Sage Publications), 18-34.

Lusher, D. and J. Koskinen and G. Robins (eds.) (in press). *Exponential Random Graph Models for Social Networks: Theories, Methods and Applications* (New York: Cambridge University Press. Structural Analysis in the Social Sciences Series).

Lusher, D. and N. Haslam (eds.) (2007). *Yearning to Breathe Free: Seeking Asylum in Australia* (Sydney: Federation Press).

Marr, D. and M. Wilkinson (eds.) (2003). *Dark Victory* (Sydney: Allen and Unwin).

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27, 1: 415-444.

Migration Amendment (Designated Unauthorised Arrivals) Bill, Parliament of Australia, Senate (2006). Available: http://www3.austlii.edu.au/au/legis/cth/bill/mauab2006521/. [May 2011]

Monge, P. R. and N. S. Contractor (2003). *Theories of Communication Networks*. New York: Oxford University Press.

Noack, A. (2007). "Energy Models for Graph Clustering." *Journal of Graph Algorithms and Applications* 11, 2: 453-480.

Palmer, M. (2005). "Inquiry into the Circumstances of the Immigration Detention of Cornelia Rau." Available: http://www.immi.gov.au/media/publications/pdf/palmer-report.pdf. [May 2011]

Park, H. W. (2003). "Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web." *Connections* 25, 1: 49-61.

Pattison, P. and G. Robins (2002). "Neighbourhood-Based Models for Social Networks." *Sociological Methodology* 32: 301-337.

Pattison, P. and S. Wasserman (1999). "Logit Models and Logistic Regressions for Social Networks; II: Multivariate Relations." *British Journal of Mathematical and Statistical Psychology* 52, 2: 169-193.

Robins, G., P. Elliott, and P. Pattison (2001a). "Network Models for Social Selection Processes." *Social Networks* 23, 1: 1-30.

Robins, G., P. Pattison, and P. Elliott (2001b). "Network Models for Social Influence Processes." *Psychometrika* 66, 2: 161-189.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007). "An Introduction to Exponential Random Graph (p*) Models for Social Networks." *Social Networks* 29, 2: 173-191.

Robins, G., P. Pattison, and P. Wang (2009). "Closure, Connectivity and Degree Distributions: Exponential Random Graph (p*) Models for Directed Social Networks." *Social Networks* 31, 2: 105-117.

Rodd, C. P. (2007). "Boats and Borders: Asylum Seekers and Elections, 1977 and 2001." In D. Lusher and N. Haslam (eds.), *Yearning to Breathe Free: Seeking Asylum in Australia* (Sydney, Australia: Federation Press).

Rogers, R. (2010). "Mapping Public Web Space with the Issuecrawler." In C. Brossard and B. Reber (eds.)*Digital Cognitive Technologies: Epistemology and Knowledge Society* (London: Wiley).

Roy Morgan Research (2005). "What Happened on November 10? Did the 'Race Card' (Border Protection) Swing the Electorate?" Available: http://www.roymorgan.com/news/polls/2001/3476/. [May 2011]

Shumate, M. and L. Dewitt 2008). "The North/South Divide in NGO Hyperlink Networks." *Journal of Computer Mediated Communication* 13, 2: 405-428.

Skvoretz, J. and K. Faust (2002). "Relations, Species, and Network Structure." *Journal of Social Structure* 3, 3. Available: http://www.cmu.edu/joss/content/articles/volume3/SkvoretzFaust.html. [May 2011]

Snijders, T. A. B. and J. Koskinen (forthcoming). "Longitudinal Models." In D. Lusher, J. Koskinen and G. Robins (eds.) *Exponential Random Graph Models for Social Networks: Theories, Methods and Applications* (New York: Cambridge University Press. Structural Analysis in the Social Sciences Series).

Snijders, T. A. B., J.H. Koskinen and M. Schweinberger (2010). "Maximum Likelihood Estimation for Social Network Dynamics." *The Annals of Applied Statistics* 4, 2: 567-588.

Snijders, T. A. B., C. E. G. Steglich, M. Schweinberger and M. Huisman (2008). *Manual for SIENA Version 3.2* (Groningen: ICS, University of Groningen; Oxford: Department of Statistics, University of Oxford). Available: http://www.stats.ox.ac.uk/~snijders/siena/s_man312.pdf. [May 2011]

Snijders, T. A. B. (2006). "Statistical Methods for Network Dynamics." In Luchini, S.R. et al. (eds.), *Proceedings of the XLIII Scientific Meeting, Italian Statistical Society (CLEUP)* Padova, Italy: 281-296.

Snijders, T. A. B., P. Pattison, G. Robins, and M. Handcock (2006). "New Specifications for Exponential Random Graph Models." *Sociological Methodology* 55, 99-153.

Snijders, T. A. B. (2001). "The Statistical Evaluation of Social Network Dynamics." *Sociological Methodology* 31, 361-395.

Steglich, C.E.G., T.A.B. Snijders, and M. Pearson (forthcoming). "Dynamic Networks and Behavior: Separating Selection from Influence." *Sociological Methodology.*

Thelwall, M., (2009). *Introduction to Webometrics: Quantitative Web Research for the Social Sciences* (San Rafael, CA: Morgan & Claypool).

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

UNHCR (2004). *UNHCR Resettlement Handbook*. Available: http://www.unhcr.org/pages/4a2ccba76.html. [May 2011]

Universal Declaration of Human Rights (1948). Available: http://www.un.org/en/documents/udhr/. [May 2011]

van Dalen, H. P. and K. Henkens (2001). "What Makes a Scientific Article Influential? The Case of Demographers." *Scientometrics* 50, 3: 455-482.

van de Donk, W., B. Loader, P. G. Nixon, and D. Rucht (2004). "Introduction: Social Movements and ICTS." In van de Donk, W., B. Loader, P. G. Nixon, and D. Rucht (eds.), *Cyberprotest: New Media, Citizens and Social Movements* (London and New York: Routledge).

Vinkler, P. (1998). "Comparative Investigation of Frequency and Strength of Motives toward Referencing: The Reference Threshold Model." *Scientometrics* 43, 1: 107-127.

Wang, P., G. Robins, and P. Pattison (2006). "LPNet: A Program for the Simulation and Estimation of Longitudinal Exponential Random Graph Models" (University of Melbourne).

Wasserman, S. and P. Pattison (1996). "Logit Models and Logistic Regressions for Social Networks: 1. An Introduction to Markov Graphs and *p*." *Psychometrika* 61, 3: 401-425.

White, H., B. Wellman, and N. Nazer (2004). "Does Citation Reflect Social Structure? Longitudinal Evidence from the 'Globenet' Interdisciplinary Research Group." *Journal of the American Society for Information Science and Technology* 55, 2: 111-126.

Footnotes

[1] Ackland and O’Neil (forthcoming) attempt to address that theoretical gap in the literature.

[2] Shumate and Dewitt (2008) also used ERGM in their analysis of the hyperlinking of NGOs, but the context was very different to that here. In particular, their primary goal was to ascertain the structure of hyperlinking between HIV/AIDS NGOs located in the South and North, in relation to theories about how the Internet is transforming spatial relations.

[3] The VOSON System is shortened to "VOSON" in this paper.

[4] A *system* is a set (or collection) of interdependent elements. In biology, a system is a set of species who are interdependent e.g. predator/prey. Key to a definition of a system is the concept of *boundaries*, which determine which elements are in the system, and which are not. A *social system* is a system where the elements are individuals and groups (or "actors") in society, and the interdependence between the actors is known as *social structure*.

[5] A *triad* is three nodes that are connected to one another, while a *transitive triad* is where each path of length 2 is closed by a tie from the start node to the end node. That is, if A links to B and B links to C, then for this triad to be transitive A must also link to C.

[6] Actor-relation effects are also sometimes referred to as *actor attributes*, but we prefer the former term as it more clearly refers to the intersection of the social ties and the attribute of the network actor.

[7] The fact that actor popularity is modeled as a structural effect highlights the fact that actor-relation effects are defined as ties that are created because of non-graph-theoretic node attributes.

[8] Also referred to as exponential families of random graphs.

[9] The following is an introduction to ERGM aimed at readers who are new to this technique. See Contractor, Wasserman, and Faust (2006) and Robins, Pattison, Kalish and Lusher (2007) for a more detailed introduction.

[10] Note that homogeneity constraints are typically used to reduce the number of parameters. If, for example, the reciprocity network motif is assumed, then there would be a reciprocity parameter for each pair of actors, leading to an unwieldy number of parameters for most networks. By constraining this parameter to be equal across all pairs of actors the model becomes easier to solve (this introduces additional error to the estimation of tie formation, but this error can be incorporated into the model as statistical noise).

[11] Note that we use the term "effect" and "parameter" interchangeably. Table 1 also includes the LPNet parameter names – these are explained further in Section 5.

[12] Negative sender effects therefore do not mean that there are fewer ties sent than expected.

[13] Hammersley and Clifford's theorem remains unpublished, but a proof was provided by Besag (1974).

[14] This distinction is further evident in the fact that with counts regressions, the unit of analysis is the actor or node while, with ERGM, the unit of analysis is the tie.

[15] Vinkler (1998) and van Dalen and Henkens (2001) also found that citations primarily reflect normatively-endorsed behavior in science. As White et al. (2004, p. 125) put it, "The evidence [regarding citation behavior] points instead toward intellectual networks ... as the real origin of intercitation. Who you know pays off only if the people you know have something worth knowing – something plainly relevant to your own claims."

[16] This of course is not strictly true, since publication delays might mean that two articles cite one another, however this would be the exception rather than the rule.

[17] However, we emphasize that this conclusion may not be relevant beyond citation networks – it is entirely possible that a *collaboration network*, where the nodes are people and the ties are collaborations between people, might involve significant levels of purely structural tie formation.

[18] Shumate and Dewitt (2008) follow Fulk et al. (1996) in extending the definition of public goods to include information and computer-mediated goods. They argue that a hyperlink network exhibits the two qualities of public goods: non-rivalry (the act of one person searching the network to locate information or resources does not preclude others from doing the same) and impossibility of exclusion (all people with a computer and Internet connection can access the hyperlink network).

[19] A web crawler is a program that automatically traverses a web site by first retrieving a web page (for example, a political party homepage) and then recursively retrieving all web pages that are referenced (e.g. following hyperlinks throughout the site).

[20] VOSON has been available for evaluation by university-based researchers and students since mid-2006, however it was only in early 2008

that data collection facilities were made generally available. See http://voson.anu.edu.au/ [May 2011] for further details.

[21] See http://nodexl.codeplex.com/ [May 2011] for information on NodeXL, and Ackland (2010) for further details on the VOSON Data Provider for NodeXL.

[22] http://socscibot.wlv.ac.uk/. [May 2011]

[23] http://www.issuecrawler.net/. [May 2011]

[24] The website of the UK's National Centre for e-Social Science (http://www.merc.ac.uk/?q=node/686) [May 2011] is a very useful resource on this topic.

[25] VOSON also uses web services to connect the various distributed research resources and hence it is possible to have different services running on servers in different administrative domains; e.g. the data collection service (web crawler) running from one university, the visualization service from another, and the SNA routines from a third. This leads to the possibility of various independent research groups running their own analytical services, thus promoting choice and variety of tools. For social-scientific research into the web, a diversity of methods (e.g. web mining, text mining, statistics) are required, and it is impossible for a single tool provider (or "one-stop shop") to provide all the necessary methods.

[26] It should be noted that the first-named author was at the time of data collection and writing a coordinator of one of the advocacy groups in this study.

[27] Even though we had determined the final "analysis" dataset contained only 144 seed sites, we re-crawled the entire 211 original seed sites.

[28] The '.com' domain is intended for commercial entities (that is, companies); '.gov' is used by government agencies; '.edu' is reserved for educational facilities; '.net' is used by many types of organizations and individuals globally, but was historically intended for use by Internet service providers; and '.org' is intended for use by the non-commercial or non-government sector. See http://www.iana.org/domains/root/db/ [May 2011] for more details.

[29] Websites are given initial random positions and the FDG algorithm then shifts the nodes in an attempt to minimize the energy that results from repulsion forces (nodes repel each other as if they were electrostatic charges) and attraction forces (hyperlinks are modeled as springs that draw together connected sites).

[30] The use of a social selection ERGM assumes that the attributes of the nodes are fixed and models the social ties. Other tools in the PNet suite are: iPNet (social influence), XPNet (multivariate), BPNet (bipartite), and LPNet (longitudinal social selection). Tools in the PNet suite use a JAVA graphical user interface (GUI) and are freely available for download from http://www.sna.unimelb.edu.au/pnet/pnet.html. [May 2011]

[31] GOF details are presented are presented in the Annex.

[32] The one exception to acceptable GOF of non-included parameters for Model B was the Global Clustering Ctm: 0.326 Mean = 0.301 (0.011) t = 2.362. Ideally, the model would not have an extreme score for this variable. Nonetheless, given it is the only extreme parameter and is a global parameter, we suggest that this is a good model for the data. Indeed, it is the best fitting model we could find for these data.