Chapter 10: Social Network Analysis (Ackland and Zhu)

in P. Halfpenny and R. Procter (2015) (eds) *Innovations in Digital Research Methods*, SAGE Publications.

Table of Contents

1. Introduction

The last decade has witnessed a surge in research into social networks using digital trace data collected from the Web. While sociologists have been studying social networks for decades the interest in researching social networks from other disciplines (e.g. other parts of the social sciences, applied physics, computer science) has arguably been triggered by the availability of data on social interactions in social network sites such as Facebook and information sharing environments such as newsgroups, blogsites, and microblogs (e.g. Twitter).

This chapter begins with an introduction to social networks analysis (SNA), outlining the major concepts and how the "network perspective" differs from other social scientific approaches to studying human behaviour. Section 3 then provides examples of network research using digital trace data, highlighting some of the methodological approaches and tools that are available for this type of research. Section 4 discusses the challenges and opportunities for network research using digital trace data. Section 5 presents some reflections and conclusions.

2. Networks and the relational perspective for studying behaviour

This section provides a brief introduction to network terminology and concepts, and introduces the network or relational perspective for studying behaviour.

2.1 Network terminology

A network is a set of nodes (vertices or entities) and a set of ties (edges or links) indicating connections or relations between the nodes.[1] In a social network, the network nodes are typically

1 For more on social network methods see Wasserman and Faust (2004) and Hanneman and Riddle (2005). For social network analysis in the context of social media data, see Hansen et al. (2010).

people and the network ties are relations between people. However, social network analysis is also used to study the behaviour of organisations and groups. A visual representation of a social network is called a sociogram, often informally referred to as a network map.

There are two types of attributes of nodes: graph-theoretic attributes are derived from the network (e.g. number of connections, or “degree”), while non-graph-theoretic attributes are intrinsic to the node (e.g. gender, race, age). Discrete or categorical node attributes (e.g. race or gender) can be represented in sociograms by the colour or shape of the node, while numerical attributes (e.g. age, degree) can be represented by the size of the node.

Edges in a network can represent different connections (collaboration, kinship, friendship, citations, transactions, shared interests, etc.). There are two major types of edges. Directed edges have a clear origin and destination (e.g. person A nominates person B as an expert in a particular field, Twitter user A follows user B) and they may or may not be reciprocated. Directed edges are represented in sociograms as a line with an arrow head indicating the direction of the tie. Undirected edges indicate a mutual relationship with no origin or destination (e.g. person A is a sibling of person B, person A is a Facebook friend of person B). By definition, undirected edges do not exist unless they are reciprocated. Edges can also have weights. An unweighted edge is where the edge either exists or not (e.g. a Facebook friendship either exists or doesn't), while a weighted edge has a value attached to it that indicates the strength of the relationship e.g. in a Twitter network a follower edge could be weighted by the number of re-tweets.

There are several types of social networks. An egocentric network (also known as an egonet, or personal network) consists of a focal node/person (“ego”) and the people he or she is connected to (“alters”). A 1.0 degree egonet doesn't show connections between alters, while a 1.5 degree egonet does show connections between alters. A 2.0 degree egonet also includes people who are 2 degrees of separation from ego i.e. friends of the alters who are not also friends of the ego. In contrast, a complete network consists of a set of nodes where all possible ties between the nodes are indicated.

Type of network can also be distinguished on the basis of mode. A unimodal network contains only one type of node or vertex, while a multimodal network contains more than one type of node. A bimodal network contains exactly two types of vertices, and an example of a bimodal network is an affiliation network consisting of people and the wiki articles they have edited. In this bimodal network, people don't connect directly with people and wiki pages don't connect directly with wiki pages. However, a bimodal affiliation network can be transformed into two separate unimodal networks e.g. wiki editor-to-wiki editor and article-to-article.

Finally, multiplex networks have multiple types of edges. Indeed, most social networks are multiplex, that is, actors share more than one type of tie. For example, two people working in a firm may have a tie that describes their working relationship (e.g. person A reports to person B) but they might also be friends outside of work and hence have another type of tie that represents their affective friendship. With regards to online networks, there are also examples of multiplex networks. With Twitter we can have three types of directed edges: following relationships, “reply to” relationships, and “mention” relationships. Often multiplex ties are reduced to a simplex tie e.g. a tie is deemed to exist if any of the multiplex ties exists.

While sociograms are useful for providing an intuitive overview of a social network, social network researchers have developed a wide range of network metrics that are used to quantitatively describe a given social network and the actors within the network. We can distinguish between node-level and network-level metrics. Some of basic node-level metrics are:

- Indegree, Outdegree - number of inbound, outbound ties (only defined for directed networks).
- Degree - number of ties (only defined for undirected network).
- Reciprocity - indicator of the extent to which there are mutual ties between actors.

- Betweenness centrality - indicator of the extent to which an individual node plays a "brokering" or "bridging" role in a network and is calculated for a given node by summing up the proportion of all the shortest pathways between the other actors in the network that pass through the node.
- Closeness centrality - indicator of the extent to which a given node has short paths to all other nodes in the graph and it is thus a reasonable measure of the extent to which the node is in the "middle" of a given network.

Some of the basic network-level metrics are:

- Network size - number of nodes in the network.
- Network density - number of network ties as proportion of the maximum possible number of network ties.
- Network inclusiveness - number of non-isolates as a proportion of the network size.
- Centralisation - a network-level property that is calculated for a given node-level property and it broadly measures the distribution of importance, power or prominence among actors in a given network i.e. the extent to which the network "revolves around" a single node or small number of nodes. The most highly centralised network is the star network, which comprises a single node (the "hub") that connects to all other nodes, but these other nodes do not connect with one another (and hence are referred to as "spokes").

2.2 Relational social science

Social network analysis (SNA) differs markedly from other social scientific approaches to studying human behaviour and outcomes. These differences are neatly captured in five fundamental and inter-related principles proposed by Wellman (1988), which together constitute the "network perspective".

First, the network perspective puts emphasis on the structure of relations, rather than the attributes of individual actors, in determining behaviour and outcomes. This can be seen clearly in the context of understanding the factors that contribute to labour market success, for example. The economic approach emphasises human capital (e.g. experience, qualifications) as being important in determining whether a person gets a job. In contrast, the network perspective regards the person's social network as the main driver of labour market success, since networks can provide opportunities (and also impose constraints) that impact on behaviour and outcomes. In particular, Granovetter (1973) emphasises the importance of "weak ties" (connections with people with whom you do not share any or many friends) as potential sources of innovative and useful information in the context of labour market outcomes.

Second, the network perspective focuses on pairs of actors (known as dyads) as the unit of analysis, rather than the actors themselves. A hallmark empirical technique used in approaches that do not take a network perspective is ordinary least squares (OLS) regression. In the labour market example above the human capital model is tested by regressing log wages on explanatory variables measuring the human capital of each individual, and in such an approach the unit of analysis is the individual. In contrast, a social networks approach to studying labour market outcomes might employ an empirical technique such as exponential random graph modelling (ERGM) where the unit of analysis is the dyad (see, e.g., Daraganova and Pattison 2013).

While a detailed introduction to ERGMs is beyond the scope of this chapter, it is useful to think of the technique as a way of deconstructing a given network into its constituent network motifs or configurations, representing different social relations such as homophily - "birds of a feather flock together", reciprocity "returning the hand of friendship", or triadic closure - "a friend of my friend is

also my friend".[2] The technique then tests whether particular configurations occur more (or less) frequently than would be expected by chance alone. Similar to standard regression techniques, the model estimation produces parameter estimates and associated standard errors, and if a particular network configuration occurs at greater or less than chance levels, we can then infer that the associated social force has had a significant role in the development of the social network.

Third, while non-relational social science assumes that observations are independent (the error term in the labour market regression example above is assumed to be independently drawn from a normal distribution), the network perspective explicitly assumes the *interdependence* of observations. Say we have three people: Sally, Jen, Andrew. Assume that Sally and Jen are friends and Jen and Andrew are friends and we want to model the probability of a tie forming between Sally and Andrew. If we assume observations (dyads) are independent, then this would be equivalent to saying that the probability of Sally and Andrew forming a friendship is the same in this situation as it would be if Jen wasn't friends with either of them. This is clearly not plausible, since it overlooks a basic aspect of social behaviour, triadic closure.

Fourth, the network perspective recognises that social networks can have both direct and indirect impacts on individual behaviour and outcomes. Simply put, the flow of information and resources between two people is not just dependent on their own relationship, but also on their relationships with everyone else. Smith and Christakis (2008) review research into social networks and health and contrast the network perspective, which explicitly models the impact of indirect or supradyadic network effects on individual health, with the more standard social support approach. With the social support approach, social network effects are operationalised as individual-level measures of how helpful or supportive social contacts are in terms of financial resources, assistance with practical tasks, information, emotional contact etc. (thus effectively involving an aggregation of the resources flowing through the dyadic ties that the individual participates in).

The principle that indirect social ties matter is also evident in Ronald Burt's work on structural holes - gaps or holes in the social structure of communication, which inhibit the flow of information between people (for a review, see Burt, 1992). Structural holes can present two types of strategic advantage, relating to brokerage and closure. With regards to the former, those people who bridge or span structural holes are more likely to gain opportunities for professional advancement because they are exposed to varying opinions, behaviours and sources of information, and may be able to combine this disparate knowledge in a way that provides productive advantage. Closure refers to the strategic advantage that is associated with not spanning a structural hole i.e. staying within a closed network where there is a high level of connectivity amongst actors (either directly or via a central person). Closed networks confer another important type of advantage to members: higher levels of trust and coordination (facilitated by high reputation costs for bad or unproductive behaviour, and increased probability that such behaviour can be detected), which can improve team effectiveness and efficiency by lowering labour and monitoring costs.

The fifth principle that underlies the network perspective is that people tend to belong to several overlapping social networks. That is, individuals tend to be members of several groups and the group boundaries are often fuzzy and hence hard to define.

3. Examples and tools

While almost all websites of user-generated content (UGC) are arguably online communities/networks of some sort, it is both informative and desirable for us to have a conceptual typology to delineate unique structural features for various online networks so that we can focus on "networking" issues rather than content or functions of the websites. For that purpose, we employ a 2-by-2 scheme involving two dimensions of network ties – directionality and manifestation – to

2 See Robins et al. (2007) for an introduction to ERGM.

help organise the rich and fast growing body of studies on online networks.

Directionality of ties simply refers to the nature of the relations (i.e., *directed* versus *undirected*) between any pair of nodes. Manifestation of ties refers to the substantiality of the relations, with *explicit* ties formed by active acts (e.g., invitation, acceptance, reference, etc.) between the nodes whereas *implicit* ties created from inferences (e.g., co-occurrence or interactions). As shown in Table 10.1, the scheme results in four distinct types of online networks, including i) explicitly undirected ties, ii) explicitly directed ties, iii) implicitly undirected ties, and iv) implicitly directed ties.

Table 10.1. Online Networks by Direction and Manifestation of ties

	Direction of Ties	
Manifestation of Ties	Undirected	Directed
Explicit	Friendship networks (e.g., Facebook, Google+, etc.)	Microblog networks (e.g., Twitter, Sina Weibo, etc.)
Implicit	Semantic networks (e.g., recommendation systems, social tagging systems etc.)	Newsgroups, blogs; WWW hyperlink networks

How much empirical research has each type of online network attracted? To gain a quick overview of the state of the art in online network research, we searched Web of Science (WoS) for studies that involve any of the 20 most popular social networking sites.[3] A total of 2,649 relevant articles are found between 2003 and 2012. Four discernible patterns emerge from the studies. First, an overwhelming majority (70%) of the studies focused on two "newcomers" of the top 20 list, i.e., Facebook and Twitter. Second, most (80%) of the studies appeared in the last three years (2010-2012), with an increasingly accelerated growth rate (Figure 10.1). Third, more than 120 research areas (as defined by WoS) were involved, ranging from social sciences, business, and humanities to science, engineering and medicine, suggesting that online social networks have become a truly multidisciplinary field of studies. Four, areas of science and technology led the studies, as shown in the top 5 most active areas including computer science (40%), engineering (13%), psychology (11%), business (11%), and communication (9%). In comparison, only 2% of the studies came from sociology, the traditional intellectual home of SNA.

A detailed review of the 2,600+ studies is obviously beyond the scope of the current chapter. We will instead highlight a few exemplars in the following sections. These studies are by no means "typical" or "representative" of the studies. On the contrary, they represent rather unconventional but rigorous efforts to online social network analysis, resulting in new, insightful, or even disruptive, knowledge of social networks. As such, they deserve our close attention.

3 http://en.wikipedia.org/wiki/List_of_social_networking_websites.

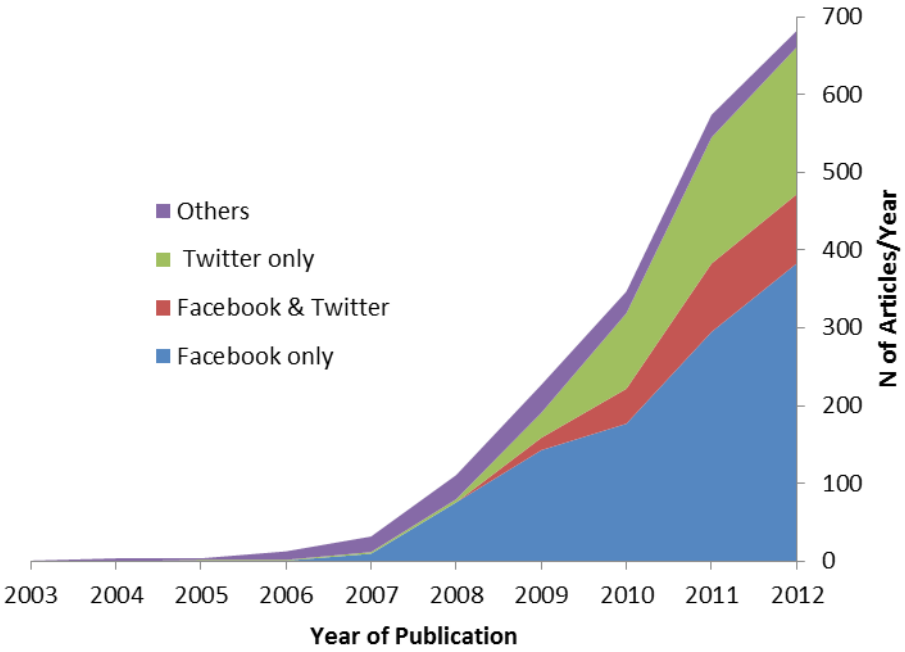

Figure 10.1. Articles on Top 20 Online Social Networks in Web of Science 2003-12

3.1 Four types of online social networks

Networks of explicitly undirected ties

These networks are the closest to the classic notion of social networks, i.e., friendships that require mutual consent to establish, interactions often closed to members only, etc. As such, many of the previous studies on online friendship networks are motivated to extend existing theories, such as formation and effects of social networks, from offline settings to online settings.

Ugander et al. (2012) present a number of challenges to our existing knowledge of social network formation. For example, they report that the probability for a non-user to join Facebook, when invited by a friend who has already used Facebook, is not affected by the popularity of the inviter (i.e., the degree of the inviter), nor by the size of contact neighbourhood (i.e., the number of common friends between the inviter and the invited). Instead, the likelihood of recruitment is affected by the number of distinct components in the contact neighbourhood.[4] Most surprisingly, the impact of component diversity on recruitment is *negative* in that the less diverse (i.e., the fewer) the components in a contact neighborhood, the more likely the non-user is to join Facebook. The findings sharply confront almost all existing relevant studies. However, given the huge size of the data (50 million users from Facebook), the rigorous design of the field experiment design, and the multiple statistical controls in the analysis, the findings certainly carry a lot of credibility.

Bond et al. (2012) carry out another experimental study using data from Facebook, which demonstrates the materialised effects of online social networks on real world behaviour. Based on a randomised online experiment design, over 60 million users were shown messages on their Facebook page with or without "social information" showing how their friends voted in the 2012 U.S. Congressional Election. The results show that those users who are exposed to the social information have a turnout rate of 0.3% higher than those who are not. Such a small difference would have been easily dismissed in any conventional experiment study. However, the big sample

4 A component is a set of nodes that are connected (either directly or indirectly) to one another.

A strongly connected component (only defined for a directed network) is where each node is reachable by all other nodes.

size of the current study, which amounts to a narrow sampling error of ±0.012%, makes the observed difference to be highly significant not only in statistical sense ($z = 24$, $p < 0.001$) but also in practical sense (180,000 users who can easily change the outcome of many local or national elections).

Networks of explicitly directed ties

The most visible examples in this type of networks are undoubtedly Twitter, Sina Weibo, and other microblog sites. They differ from Facebook and other friendship networks primarily in their one-way, public mode of the relations among users. Most microblog sites were originally intended to be used as friendship networks similar to Facebook but unexpectedly, users "reinvented" the media function of the sites and transformed them into the news sharing platforms that we know today. This unprecedented chapter of social network history is timely recognised and documented by Kwak et al. (2010) and Wu et al. (2011), among several other similar studies.

Kwak et al. (2010) start with a simple but unsettled question of Twitter at the time: Is it a social (i.e., friendship) network or a news medium? They focus on a key property of social network – the reciprocity between followers and followees – on Twitter. As it turns out, less than one quarter (22%) of the user pairs reciprocally follow each other whereas more than three quarters (78%) are of one-way relationship. Coupled with other topological features of Twitter, such as a short effective diameter and a non-Power Law distribution of indegrees, it becomes clear that the microblog site is not a typical social friendship network but a news forwarding network.

Wu et al. (2011) move further to examine exactly how Twitter functions as a new medium – i.e., who says what to whom – by following the classic 5W model of mass communication (Lasswell, 1948). They first searched for the prevalence of four pre-determined categories of news sources (or elite users), including celebrities, news media, institutions, and professional bloggers, who account for only half a million (or 1.2%) of the user population on Twitter. The most influential elite users (accounting for 0.05% of the population) attract 50% of the following links. The highly uneven distribution of attention illustrates the star-topology of microblog networks.

Networks of implicitly undirected ties

Network of implicit ties are formed based on "hidden" connections. Of implicit networks, those with undirected ties are the most "invisible" because the links among the nodes are constructed or inferred by social network analysts post hoc, based on semantic similarity (e.g., co-usage or co-occurrence of keywords, tags, favourites, acts, etc.) between nodes of pairs. While the validity of such networks is sometimes questionable, they are often more informative about the substantive content of social interactions than what link-based social networks could reveal. Probably because of the latent nature of semantic similarity, it has not been frequently used in online social network research. However, a few existing studies demonstrate the unique value of such an approach.

For example, Capocci et al. (2010) evaluate the observed semantic similarity of photo tags between Flickr users against the corresponding friendship between the same users. The results show that semantic similarity does not result from social interactions among linked friends (i.e., peer influence hypothesis), but from existing similar background of users (i.e., social selection or homophily hypothesis).

Networks of implicitly directed ties

Newsgroups (including various bulletin board systems (BBSs) and other online forums) are the most representative examples of implicitly directed ties. On the first glance, newsgroups do not look like a typical social network because nodes and edges are not directly observable. As such, researchers have to first "explicate" nodes and edges of newsgroups before carrying out formal network analysis. Participants in BBS/forum discussions are often taken as nodes and their opinion

exchanges as edges. Since online opinion exchanges usually involve a sequence of question-reply or post-comment, the derived networks are directed by nature. Occasionally, threads (i.e., topics) of exchanges are used as nodes whereas participants in the same threads are treated as edges between topical nodes, which necessarily result undirected networks.

An example of research using newsgroup data is Welser et al. (2007), who investigate the use of network techniques for identifying "answer people" - individuals who mainly respond to questions posed by other users instead of posing their own questions, or getting involved in unproductive "flame wars". The authors argue that being able to identify answer people is important for developing better approaches for cultivating online communities of practice where such sharing of information is prevalent. They find that answer people tend to participate in threads initiated by others and typically only contribute one or two messages per thread. Second, the ego networks of answer people tend to contain alters who themselves answer few, if any, questions posed by others. Further, the 1.5-degree ego network of answer people tends to have small proportions of triads (i.e. their neighbours are not neighbours of each other) and they have few intense ties (i.e. they seldom send multiple messages to the same recipient).

While all social networks are built on hyperlinks, the term "hyperlink networks" refer to a specific genre: graphs of webpages or websites that are tied together with hyperlinks. In a seminal study, Broder et al. (2000) examine the topological structure of the global WWW based on 200 million webpages collected by search engine AltaVista. They conclude that the Web was organised as a "Bow Tie" consisting of four equally-sized components, with a strong connected component (SCC) in the center, an IN component with one-way links to SCC, an OUT component with one-way links from SCC, and a component of Tendrils isolated from SCC. Later studies of hyperlinks at the national scope reveal a much more centralised graph, which looks like a "Daisy" with SCC accounting for 80% of the webpages in Italy, UK, and five Indonesian nations (Donato et al., 2005) or a "Teapot" with 44% of the webpages in China falling into SCC (Zhu et al., 2008). Parallel to these graph models of hyperlink networks by computer scientists, Halavais (2000) identifies the existence of national borders on the Web, based on the pattern that websites across a dozen nations were more likely to link domestic websites than foreign websites.

Hyperlink networks also prove to be insightful in examining the interplay between social movement actors. For example Ackland and O'Neil (2011) study environmental activist hyperlink networks, extending Diani's (1992) network-theoretic approach to studying social movements to the online world. In keeping with our above classification of hyperlink networks as networks of implicitly directed ties, the authors do not contend that hyperlinks between environmental activist websites necessarily reflect the exchange of real-world resources. Instead they emphasise the role of hyperlinks in the exchange of symbolic resources, thus helping to establish boundaries of inclusion/exclusion, and hence the formation of collective identity.

3.2 Tools for collecting and analysing online network data

This section presents a brief summary of tools for collecting and analysing online network data.

Tools for collection of online network data

Table 10.2 revisits the typology for online networks introduced above, presenting an example of a tool for collecting data from each type of network. It should be noted that the table only shows data collection tools that are publicly available, and only one example tool is shown for each type of network.

Table 10.2. Tools for Collection of Online Network Data

	Direction of Ties

Manifestation of Ties	Undirected	Directed
Explicit	Social Network Importer for Facebook networks (http://socialnetimporter.codeplex.com)	Tweepy Python library for Twitter API (http://tweepy.github.io)
Implicit	Python Flickr API kit (http://stuvel.eu/flickrapi)	VOSON - hyperlink network collection and analysis (http://voson.anu.edu.au)

Social Network Importer is a plugin for NodeXL[5], which is a free Excel 2007/2010 template for analysing networks in the familiar Excel spreadsheet environment. The predecessor of Social Network Importer was NameGen[6], which was designed by Bernie Hogan at the Oxford Internet Institute (see, e.g., Hogan 2010). Social Network Importer queries the Facebook Application Programming Interface (API), allowing the extraction of ego network data for a given Facebook user. Depending on account privacy settings for ego and alters, the tool will also collect Facebook profile data and return the 1.5 degree ego network. According to the Facebook API terms and conditions, the data can only be collected for an ego who has provided his or her Facebook username and password, and hence Social Network Importer is currently mainly useful for researchers who want to collect their own ego network data or that of a small number of participants (who would need to use NodeXL on a machine that the researcher has access to). In contrast, NameGen was available as a Facebook application and it allowed the creators of NameGen to collect ego network data for people who consented to participate in the study (where consent was granted via the installation and use of the NameGen Facebook application).

While Social Network Importer conveniently hides the interactions with the Facebook API from the researcher, the *Tweepy Python library for Twitter API* is much more low-level in that its use requires the researcher to be able to program in Python. Typical use of Tweepy might involve the researcher querying the Twitter Search API to find all recent tweets that contain a particular hashtag. The Twitter User API could then be used to collect the directed follower network of the authors of those tweets.

Similar to Tweepy, the *Python Flickr API kit* is designed for Python programmers who want to programmatically interact with the Flickr photo sharing website. A research use of the Python Flickr API kit may involve getting a list of meta data (e.g. descriptive tags) on photos uploaded by a particular Flickr member (or members) and then iterating over the list of meta data and constructing a semantic network where an undirected and weighted tie between two tags indicates the number of times they were jointly used to describe a particular photo.

Finally, the VOSON tool for hyperlink network collection and analysis (e.g. Ackland, 2010) is available as both a web application and a plugin to NodeXL.[7] Users can enter a list of seed URLs (typically, entry pages to websites) and the web crawler will then crawl through each site and collect outbound hyperlinks and text content. Optionally, the crawler will also return inbound hyperlinks to each page in the site (this is currently achieved via the VOSON software accessing the Blekko web search engine API[8]). VOSON allows the user to construct networks of web pages or websites, and these can be visualised in the web application and it is possible to download networks for analysis in other network analysis tools.

5 http://nodexl.codeplex.com

6 See http://people.oii.ox.ac.uk/hogan/software/

7 The first-named author created the VOSON software and has a financial interest in its commercialisation.

8 http://blekko.com

Tools for analysis and visualisation of networks

Having collected online network data using, for example, one of the tools mentioned above, the researcher then needs to decide what software will be used for network analysis. The following is a brief and non-exhaustive list of freely-available software for network analysis.

- NodeXL (http://nodexl.codeplex.com) was mentioned above in the context of data collection, but it also provides a menu-driven environment for network visualisation and analysis.
- Pajek (http://pajek.imfm.si/doku.php) is a Windows-based menu-driven package, known for its ability to handle large networks.
- Statnet (http://statnet.csde.washington.edu/) is a suite of R (open source statistical software) libraries for network handling and analysis, including ERGM.
- NetworkX (http://networkx.github.io/) NetworkX is a Python language software package for network analysis.
- igraph (http://igraph.sourceforge.net/) is a library for network analysis that runs in both R and Python.
- Gephi (https://gephi.org/) runs on Windows, Linux and Mac OS and is a menu-driven network visualisation tool.
- PNet (http://sna.unimelb.edu.au/) is a menu-driven Windows package for ERGM.
- UCInet (https://sites.google.com/site/ucinetsoftware/home) is a menu-driven Windows package for social network analysis.

4. Challenges and opportunities

This section identifies challenges and opportunities for SNA researchers working with digital trace network data (network data that are collected unobtrusively from the Web).

Social network or information sharing network?

Is it always reasonable to conceive of an online network as exhibiting the hallmarks of social networks (e.g. interdependence, evolution, one-to-one ties, multiplex ties)? That is, having outlined the network perspective above, it is prudent to ask whether a given online network is likely to exhibit properties such that it can be viably studied under this perspective using SNA techniques.

Referring back to the typology of online networks introduced in Table 10.1, it was noted above that online networks involving explicitly undirected ties (e.g. Facebook) are more easily conceived of as social networks and Wimmer and Lewis (2010), for example, have used ERGM to provide new insights into friendship homophily using Facebook data.

In contrast, it is questionable as to whether modelling microblogs such as Twitter (networks of explicitly directed ties) as social networks is valid. The founders of Twitter have emphasised that it is an information sharing network, not a social network, and as noted above, research by Kwak et al. (2010) shows that Twitter resembles more closely an information network, compared to a social network.

The online network that is probably the hardest to conceive of as a social network is a semantic network (network of implicitly undirected ties). However, it is of interest to note that Capocci et al. (2010) made use of semantic networks in their study of role of social influence (people becoming more like their friends) versus social selection (similar people becoming friends) in determining the photo tagging behaviour of Flickr members, and their findings (discussed above) are largely

consistent with what Lewis et al. (2012) find in their extensive analysis of Facebook data. In other words, networks inferred from semantic similarity provide an effective and efficient way to address some long-standing debates in social network research, such as the causal direction between social selection and peer influence, although Capocci et al. (2010) themselves do not acknowledge such implications.

Finally, while it is hard to conceive of all Web 1.0 hyperlink networks (networks of implicitly directed ties) as social networks, Lusher and Ackland (2011) have argued that the network perspective is valid for Web 1.0 hyperlink networks comprising particular types of actors. ERGM has been used by several authors to model NGO hyperlinking (e.g. Shumate and Dewitt 2008, Gonzalez-Bailon 2009, Ackland and O'Neil 2011, Lusher and Ackland 2011). In contrast, the hyperlinking behaviour of government agencies and academic research teams is possibly less likely to conform to the network perspective and consequently a lot of studies of these types of Web 1.0 actors have involved techniques that do not emphasise relational structure. In particular, webometric analysis of government visibility or centrality on the Web (Escher et al. 2006) and the academic authority of research teams in the biotech sector (Barjak and Thelwall 2008) have eschewed the construction of complete network data and instead reduce network structure to being an attribute of the actors (indegree or outdegree) that, in the case of Barjak and Thelwall (2008), is then used as a dependent variable in regression analysis.

Construct validity of online network data

While researchers from some fields (e.g. new media and virtual ethnography) study online behaviour in its own right, other social scientists are often interested in online network data because they can potentially tell us something new about offline social processes, such as friendship formation, partnering behaviour, dynamics of teams and organisations, and social influence. The promise of online environments such as Facebook, Twitter and Second Life is that they can provide large quantities of precisely timestamped social behavioural data that would be impossible to collect using obtrusive methods, such as interviews or surveys.

However, many sociologists in the SNA tradition believe that the viability of online network data in this context is predicated on our being able to sufficiently demonstrate that there is a "mapping" (Williams 2010) between the online and offline world, or that online data have "construct validity" (Burt 2011) with regards to important offline behaviour. Burt (2011, p. ?) argues that "... the advantages of network data in virtual worlds are worthless without calibrating the analogy between real and inworld. If social networks in virtual worlds operate by unique processes unrelated to networks in the real world, then the scale and precision of data available on social networks in virtual worlds has no value for understanding relations in the real world."

So, a major challenge for social scientists working with online network data is establishing that the data have construct validity for the purpose at hand. There are three ways in which construct validity might be demonstrated (Ackland 2013).

First, the construct validity of web data may be demonstrated by showing that the online network displays structural signatures that are consistent with those displayed by real-world actors. For example: does Facebook friendship network data display homophily on the basis of race and ethnicity (Wimmer and Lewis 2010)? Are divisions between different groups in the environmental social movement evident in hyperlink networks (Ackland and O'Neil 2011)? To what extent is political affiliation reflected in political blog networks (Adamic and Glance 2005)? If web networks display structural signatures that are significantly different to those shown in the real world, then the construct validity of the data in the real world is questionable.

Second, it may be possible to assess the construct validity of web data by testing whether variables constructed from web data are correlated with other accepted measures of the construct. For example, if counts of inbound hyperlinks to academic project websites are correlated with other characteristics of academic teams (e.g. publications, industry connections) that are used as proxies

of academic authority/performance, then this is evidence of the construct validity of hyperlink data in the context of scientometrics, which is the science of measuring and analysing science (see, e.g. Barjak and Thelwall 2008).

Finally, the construct validity of web data may be demonstrated if it can be shown that an actor's position in an online network influences his or her performance in a manner that accords with what is found in the real world. For example, Burt (2011) studies brokerage and closure in Second Life, making use of three types of relational data: one-to-one friendships (which operate in a manner similar to Facebook), group membership (again, similar to that in Facebook) and rights granted to friends. With regards to the last, Second Life allows a user to grant another user three levels of access (which imply increasing levels of trust): the right to know when you are online, the right to know your location in Second Life, and the right to directly modify your online inventory (thus impacting on the appearance and behavior of your avatar).

Burt further constructs ego networks by defining and quantifying four types of directed tie: "no tie" – users i and j are not friends; "weak tie" – users i and j are friends but i either granted j no rights or only the right to know when j is online; "average tie" – i and j are friends, and i granted j the right to know i's location inworld; "strong tie" – i and j are friends, and i granted j the right to modify i's online inventory.

With regards to the closure hypothesis, there is no obvious Second Life analog to team effectiveness and efficiency measures and so Burt focuses on the intermediate variable of trust and tests the hypothesis that closure is associated with higher levels of trust in relationships. Trust is measured by the three levels of rights that friends can grant each other and Burt finds support for the closure hypothesis, showing that the level of trust between two users is positively related to the level of network closure around the friendship, measured by the number of indirect connections between ego and the friend.

With regards to brokerage, the hypothesis is that occupying structural holes in Second Life is correlated with achievement, measured as the contribution of ego to the formation and maintenance of groups (one of the main types of infrastructure that attracts people to Second Life). It is found that people who have greater access to structural holes (measured by the number of non-redundant friends i.e. friends who are not connected to each other) maintain more groups and their groups are more likely to be successful in terms of membership and activity.

The above arguments draw on the conventional definition of construct validity. As online networks keep becoming increasingly ubiquitous, they may actually have started to evolve into a new "world" that is fundamentally different from the real world. The scenario certainly challenges the requirement that the construct validity of online networks lies in the consistency between offline and online worlds.

Causality versus correlation

Identifying causality is one of the major bugbears of empirical social science. The problem arises because social scientists are often trying to identify the impact of a variable that the individual has some degree of choice over (e.g. the impact of years of schooling on wages, impact of peers on student performance). In the standard human capital model introduced above, for example, it has long been recognised that OLS estimates of the impact of an additional year of education on wages are biased upwards because years of schooling are likely to be correlated with unobservable variables such as ability (more able students are likely to get more years of schooling, on average).

The growth of interest over the past decade in the role of social networks in behaviour and outcomes has only served to highlight the methodological challenges associated with identifying causality. The problem for social scientists is that, unlike their counterparts in the natural sciences, social science data are typically observational rather than experimental, and it is therefore very difficult to control the variable of interest. However, the Web offers two major opportunities for

overcoming methodological limitations relating to discerning the causal impact of social networks.

First, a major advantage of Web network research is the potential for collecting rich and precise time-stamped data. Burt (2011) noted that his results are based on a single snapshot of activity within Second Life; the fact that the sequence of friendship and group formation is not taken into account prevents any conclusions about causation. In particular, it may be that two people with high level of trust in their relationship subsequently acquire many mutual friends (hence building network closure). Similarly, a person who has established successful groups in Second Life may acquire friendships from diverse parts of the virtual world, with access to structural holes thus resulting from achievement, rather than the other way around. However, Burt notes that the availability of precisely timestamped network data from online environments such as Second Life may allow researchers to better understand the causal relationship between network position and outcomes. Although Burt does not explicitly explain how this will work, we believe that the timestamped data (usually on a daily or finer scale) allow social network analysts to test a variety of alternative models between pairs of reciprocal effects, something always difficult or impossible to do with offline data of crude timestamp (usually on a yearly or larger scale).

Second, the Web allows for field experimentation (experiments that are conducted outside of the laboratory but where the researcher still has control over the variable of interest) that would be very difficult to conduct in the real world and better equips researchers to discern the causal impact of social networks.

Centola (2010) uses an online field experiment to study the social transmission of health behaviour in an attempt to uncover the exact processes by which people are influenced by their social network. The "strength of weak ties" literature (e.g. Granovetter 1973) suggests that the diffusion of an innovation (e.g. information, behaviour) through a network is going to occur more efficiently where there is low redundancy of ties i.e. your neighbours do not tend to be connected to one another, and hence do not provide the same information. However, Centola and Macy (2007) suggest that while the "weak ties" argument might hold for innovations such as information or disease, it may not hold up for transmission of behaviours where social affirmation from multiple sources is required for successful transmission (transmission will be aided by network structures that exhibit clustering, which increases the likelihood of contact from multiple sources). Centola and Macy (2007) term this "complex contagion" and provide examples such as spread of high-risk social movements and avant garde fashions.

Centola (2010) explores the role of network structure in the promotion of social influence by creating an Internet-based health community containing 1528 participants recruited from online health communities of interest. Participants were randomly assigned health "buddies" from whom they could gain information about a new online health forum. The act of joining the health forum was the health behaviour under examination, and the study aimed to quantify the impact of number of sources of information on the probability of adopting the behaviour. Participants were randomly assigned to networks with differing levels of clustering, which meant there was variation in the number of network neighbours providing information about the new health behaviour. The author finds that the probability of adoption increased markedly when participants received social reinforcement from multiple network neighbours.

Natural research instrument

Online network data are being generated in what is effectively a huge natural research instrument, the Web. Wimmer and Lewis (2010) note that the fact that Facebook is a natural research instrument provides both opportunities and challenges for research into friendship formation.

The advantages are that there is no interviewer effect or recall error and there is no need to reduce the scope of the study (for cost reasons). This allows Wimmer and Lewis (2010) to collect complete network data for an entire cohort of undergraduate students in a residential college in the US, which would have been prohibitively expensive to do using traditional offline methods. The

fact that Facebook users provide rich "cultural preferences" data (e.g. favourite books, movies, music) also provides an opportunity for researchers. Such data would be difficult and expensive to collect using offline methods (friendship homophily research traditionally has involved only very basic demographic profile data such as age, gender and race). The availability of cultural preferences data allowed Wimmer and Lewis (2010) to provide new insights into how shared cultural tastes may influence friendship formation.

But Wimmer and Lewis (2010) also note that Facebook data provide challenges for friendship homophily researchers. In particular, what is the exact social meaning of a “facebook friend”? Further, the "cultural preferences" data (e.g. books, movies, music) reflect both true preferences and strategic presentation of self, and this may pose problems (but of course, strategic self-presentation can affect other types of social research data, however generated).

Another important challenge relating to the Web as a natural research instrument is that we only get partial information about network participants. This challenge is particularly evident in research into social influence in social media. While an individual may be influenced via interactions in social media to change his or her behaviour (e.g. attitudes to politics or consumer preferences) we are not going to know this unless the behavioural change results in a digital trace that is picked up by the researcher. The study by Aral et al. (2009) into social influence in an instant messaging network focuses on a measurable outcome (product adoption). But what about changes in a person’s political identity or willingness to vote? Unless this change is somehow (easily) identifiable in the data, or else we conduct a survey or experiment to gauge how attitudes have changed, this will never be known by the researcher.

In Twitter, how do we know when someone has been influenced, that is, what is the appropriate measurable outcome? In research into social influence in Twitter, influence has often been conceptualised as attention, for example number of followers (e.g. Kwak et al. 2010, Cha et al. 2010). That is, what is really a dynamic process (e.g., A interacts with B and as a result of the interaction, A’s behaviour changes) gets reduced to a static question of who is the most central or visible or important person in the social media network i.e. who is garnering the most attention? A current popular measure of influence in Twitter is the number of retweets (e.g. Kwak et al. 2010, Cha et al. 2010); while a retweet is a clear indication that someone has made a conscious decision to pass on information, a problem is that it is difficult to disentangle the influence of the content (some tweets are going to contain information that is innately "viral") from the influence of the original tweeter (some tweeters might have greater authority that means people tend to retweet them, regardless of the tweet content). Harrigan et al. (2012) get around this problem by using conditional logistic regression which allows the probability of retweeting to be estimated only for those people who follow the same tweeters (and hence the virality of tweets and the influence of their authors is controlled for).

Big Data and network sampling

Digital network datasets can be potentially huge. While the scale of online networks may not pose a problem in empirical techniques used by applied physicists and computer scientists, some established statistical SNA techniques (such as ERGM) currently do not scale well (there can be problems with model convergence with large and dense networks). An additional problem for social scientists is that the research often requires that more is known about the network participants than can be gathered via automatic data collection methods. This often means that only a subset of the available data are analysed, in order to make the human-coding of actors feasible. For example, in their analysis of information flows on Twitter during the 2011 Egyptian and Tunisian revolutions, Lotan et al. (2011) only analyse the 10% largest tweet flows, resulting in a subset of 963 users who either were first to post in a flow, or were retweeted or mentioned at least 15 times.

Social scientists are well versed in sampling approaches, with much empirical social science being based on representative samples of individuals or households drawn from a larger population.

However, network sampling poses a different set of challenges: even if one is able to define the population of interest in an online network, random sampling is unlikely to be appropriate since the units of observation and hence error terms are not going to be independently and identically distributed, and the sampling approach should reflect this.

Sampling of online social networks involves in general four strategies: sampling of nodes, ego-networks, sub-networks, and full networks.[9] Of the four, it is the easiest to apply probability sampling to ego-networks because their alters (i.e., neighbours) are usually directly accessible from the chosen egos.

Probability sampling of nodes is possible, although costly, from the social networking sites with a known sampling frame (e.g., the URLs of all members are directly or indirectly known). For example, many social networking sites systematically assign a numeric ID to map the home page of their users, which provide necessary information to construct a sampling frame for those sites. We have devised a random digit search (RDS) method, following random digit dialing (RDD) method used in offline telephone surveys, to detect the boundary and coverage rate of valid user IDs and then draw probability samples from the known range of URLs of a popular blog site in China (Zhu et al., 2011). It is necessary to note that the resulting samples, while providing unbiased and precise estimates of individual nodes, do not represent topological characteristics of the underlying networks from which the nodes are drawn.

Sampling of sub-networks involves the same methodological issues as those in sampling of full networks (as discussed next), plus the conceptual problem of generalisability of sub-networks. While popular among the previous studies, how universally applicable are the findings observed from student Facebook users at a single university or a few selected universities? Obviously, the issue goes beyond the scope of the current discussion.

Sampling of full networks is the ideal method for study of network characteristics. However, despite various efforts, no perfect solution has yet emerged. To begin with, as noted above, probability sampling is not only confined to certain sites, but is also incapable of capturing network topology. As such, most studies of network sampling employ various methods of snowballing. The most common, but also most problematic, method is breadth-first search (BFS), which samples every neighbour of chosen seeds and, if necessary, every neighbour of the chosen neighbors successively in the onward path until a desirable sample size is reached (Ahn et al., 2007). While highly efficient, BFS samples over-represent popular nodes, given the power-law distribution of edges on virtually all social networks (Kurant, Markopoulou, and Thiran, 2010; Zhu et al., 2013).

To overcome the systematic biasedness of BFS, several schemes of random walk (RW) sampling have been proposed. Similar to BFS, RW schemes sample successfully neighbours of chosen seeds, by which they are also snowballing (i.e., non-probabilistic, despite the name of “random walk”). Instead of including everyone along the path, RW sampling selects only one or a small number of nodes per step, based on different weighting strategies adopted by different RW schemes such as Respondent-Driven Sampling (RDS; Wejnert and Heckathorn, 2008), Re-Weighted Random Walk (RWRW; Gjoka et al., 2010), Metropolis-Hasting Random Walk (MHRW; Gjoka et al., 2010; Stutzbach et al., 2006), and Self-Adjustable Random Walk (SARW; Zhu et al., 2013). Each of the schemes presents some significant improvement over BFS. However, none of them can provide simultaneously unbiased estimates of multiple key parameters of network structure (e.g., mean degree, clustering coefficient, assortativity, etc.), when tested against large-scale, real social network data (Zhu et al., 2013). Sampling of online social networks remains one of the top challenges in the age of Big Data.

---

9 Sampling of nodes is different from sampling of ego-, sub- or full networks as the former refers to studies that focus exclusively on characteristics of individual nodes without any concerns about characteristics of the networks in which the nodes embedded. Strictly speaking, such studies of nodes are not of social network analysis. However, given the popularity of this approach in the current literature and the inherent connection between nodes and networks, we include it here.

Proprietary data

One of the advantages of researching Web 1.0 is that the data are generally available. The data may not necessarily be easy to collect (given the need for a web crawler and further pre-processing of the network data) but as long as the webmaster wants the site to be indexed by Google, then the researcher can also get the data via a web crawler. Furthermore, search companies such as Google and Yahoo also play a positive role in enabling network research by providing APIs that allow researchers to programmatically extract large quantities of hyperlink network data from search engine indexes (see, e.g., Thelwall 2004).

With the advent of Web 2.0, the opportunities for social research using web data have expanded since more people are interacting online by, for example, blogging, using Facebook and Twitter. There are more social scientists who want to use web data for social network research, but these data are actually becoming harder to obtain. While there are tools for extracting egonets from Facebook (see, e.g. Hogan 2010), research involving complete network Facebook data (e.g. that of Wimmer and Lewis 2010) is generally going to require a data sharing agreement with Facebook, which are very hard to obtain. Similarly, Burt's (2011) research into Second Life involves use of a proprietary dataset. Twitter have been very open to research and there are several tools that can be used for extracting Twitter network data (e.g. Barash and Golder 2010), but there are indications that the Twitter API will not always remain open for access by researchers.[10]

The challenge associated with proprietary data is that it might lead to proprietary researchers i.e. researchers working on topics that are most interesting to the data providers, and producing research that cannot easily be replicated or validated.

5. Reflections and conclusion

Big data of online networks have brought both opportunities and challenges to SNA researchers. The following are probably among the most important issues for SNA researchers to fully benefit from the opportunities and meet the challenges,

- Engage in collaboration with scholars from sciences and engineering. SNA researchers should take the initiatives to reach out;
- Focus on identify both similarities and differences in network structure and behaviour between the real and virtual worlds;
- Work out, preferably through interdisciplinary collaboration, scalable versions of the rich, existing models and algorithms of SNA to deal with large-scale online networks.

References

Ackland, R. (2010). "WWW Hyperlink Networks," Chapter 12 in D. Hansen, B. Shneiderman and M. Smith (eds), *Analyzing Social Media Networks with NodeXL: Insights from a connected world.* Morgan-Kaufmann, Burlington, MA.

Ackland, R. (2013). *Web Social Science*. SAGE Publications, London.

Ackland, R. and O'Neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks*, 33:177–190.

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

10 For example, in June 2013 Twitter "whitelist accounts" (which provided substantially higher number of API calls per hour than the standard account) were turned off.

Ahn, Y. Y., Han, S., Kwak, H., Moon, S. and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. *The 16th international conference on World Wide Web (WWW2007)*, Alberta, Canada.

Aral, S. and Walker, D. (2010). Creating social contagion through viral product design: A randomized trial of peer influence in networks. New York University Stern School of Business Working Paper.

Barash, V. and Golder, S. (2010). "Twitter: conversation, entertainment and information, All in one network!," Chapter 10 from D. Hansen, B. Shneiderman and M. Smith (eds), *Analyzing Social Media Networks with NodeXL: Insights from a connected world*. Morgan-Kaufmann, Burlington, MA.

Barjak, F. and Thelwall, M. (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):628–43.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489, 295–298.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33(1-6), 309-320.

Burt, R. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, Mass.

Burt, R. (2011). Structural holes in virtual worlds. Booth School of Business (Univ. of Chicago) working paper.

Capocci, A., Baldassarri, A., Servedio, V. D. P., Loreto, V., and Gualino, V. (2010). Friendship, collaboration and semantics in Flickr: From social interaction to semantic similarity. *MSM'10*, Toronto, Canada.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329:1194-1197.

Centola, D. and Macy, M. W. (2007). Complex contagion and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734.

Cha, M., Haddadi, H., Benevenuto, F., and Gummad, K. P. (2010). Measuring user influence on twitter: The million follower fallacy. 4th International AAAI Conference on Weblogs and Social Media, Washington, DC.

Daraganova, G. and Pattison, P. (2013). Autologistic actor attribute model analysis of unemployment: Dual importance of who you know and where you live. In D. Lusher, J. Koskinen and G. Robins, editors, *Exponential Random Graph Models for Social Networks*, Cambridge University Press, New York.

Diani, M. (1992). The concept of social movement. *Sociological Review*, 40(1):1–25.

Donato, D. Leonardi, S., Millozzi, S., and Tsaparas, P. (2005). Mining the inner structure of the Web graph. *The 8th international workshop on the Web and Databases (WebDB 2005)*, Baltimore, USA.

Escher, T., Margetts, H., Petricek, V., and Cox, I. (2006). Governing from the centre? comparing the nodality of digital governments. Paper presented at the 2006 Annual Meeting of the American Political Science Association, Chicago, August 31 September 4, 2006.

Gilad L., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., and boyd, d. (2011). "The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions." *International Journal of Communications*, 5:1375–1405. http://ijoc.org/ojs/index.php/ijoc/article/view/1246/613

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010). Walking in Facebook: Unbiased sampling of OSNs. *The Proceedings of IEEE INFOCOM*. San Diego, USA.

Gonzalez-Bailon, S. (2009). Opening the black box of link formation: Social factors underlying the structure of the web. *Social Networks*, 31:271-280.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–80.

Halavais, A. (2000). National borders on the World Wide Web. *New Media & Society*, 2(1), 7-28.

Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods. Riverside, CA: University of California, Riverside. Published in digital form at http://faculty.ucr.edu/~hanneman/.

Hansen, D. L., Shneiderman, B., and Smith, M. A. (2010). *Analyzing Social Media Networks with NodeXL: Insights from a connected world*. Morgan-Kaufmann, Burlington, MA.

Harrigan,N., Achananuparp, P and Lim, E-P. (2012). "Influentials, novelty and social contagion: The viral power of average friends, close communities and old news," *Social Networks*, 34(4), 470-480.

Hogan, B. (2010). Visualizing and interpreting facebook networks. In Hansen, D. L., Shneiderman, B., and Smith, M. A., editors, *Analyzing Social Media Networks with NodeXL: Insights from a connected world*. Morgan-Kaufmann, Burlington, MA.

Kwak, H. W., Lee, C. H., Park, H. S., and Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World Wide Web (WWW 2010)*, 591-600.

Kurant, M., Markopoulou, A., and Thiran, P. (2010). On the bias of BFS (Breadth First Search). *The annual conference of International Teletraffic Congress (ITC 22)*, Amsterdam, the Netherlands.

Lasswell, H. D. (1948). The structure and function of communication in society. In L. Bryston (Ed.), *The communication of ideas*. Harper and Brothers, New York.

Lewis, K., Gonzalez, M., and Kaufman, J. (2012). Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences of the United States of America, 109, 68-72.

Lusher, D. and Ackland, R. (2011). A relational hyperlink analysis of an online social movement. *Journal of Social Structure*, 12(5). Available at: http://www.cmu.edu/joss/content/articles/volume12/Lusher/.

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

Shumate, M. and Dewitt, L. (2008). The north/south divide in ngo hyperlink networks. *Journal of Computer Mediated Communication*, 13:405–428.

Smith, K. and Christakis, N.A. (2008) Social Networks and Health, *Annual Review of Sociology*, 34, 405-29.

Stutzbach, D., Rejaie, R., Duffield, N., Sen, S., and Willinger, W. (2006). On unbiased sampling for unstructured peer-to-peer networks. *Proceedings of Internet Measurement Conference (IMC 2006)*. Rio de Janeriro, Brazil.

Thelwall, M. (2004). Link Analysis: An Information Science Approach. Academic Press. Available at: http://linkanalysis.wlv.ac.uk.

Ugander, J., Backstrom, L., Marlow, C., and Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of National Academy of Science*, 109(16), 5962-5966.

Wasserman, S. and Faust, K. (2004). *Social network analysis*. Cambridge University Press, Cambridge, UK.

Wejnert, C., and Heckathorn, D. D. (2008). Web-based network sampling: efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research*, 37(1), 105-134.

Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance. In B. Wellman and S. Berkowitz (Eds.), *Social structures: A network approach* (pp. 19-61). Cambridge University Press, Cambridge, UK.

Welser, H., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. Journal of Social Structure, 8(2). Available at: http://www.cmu.edu/joss/content/articles/volume8/Welser/.

Williams, D. (2010). The mapping principle, and a research framework for virtual worlds. *Communication Theory*, 20(4):451-470.

Wimmer, A. and Lewis, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology*, 116(2):583–642.

Wu, S., Hofman, J., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on Twitter. WWW 2011, March 28-April 1, 2011, Hyderabad, India.

Zhu, J. J. H., Meng, T., Xie, Z. M., Li, G., and Li, X. M. (2008). A Teapot graph and its hierarchical structure of the Chinese Web. *Proceedings of the 17th international conference on World Wide Web (WWW'08)*, 1133-1134.

Zhu, J. J. H., Mo, Q., Wang, F., and Lu, H. (2011). A random digit search (RDS) method for sampling of blogs and other web content. *Social Science Computer Review*, 29(3), 327-339

Zhu, J. J. H., Xu, X. K., Zhang, L., and Peng, T. Q. (2013). A flexible sampling method for large-scale online social networks: Self-adjustable random walk (SARW). *The annual conference of International Network for Social Network Analysis (Sunbelt2013)*, Hamburg, Germany.