Chapter 1: Political Homophily on the Web

Robert Ackland and Jamsheed Shorish


Preprint: April 2013

1 INTRODUCTION

Homophily is a central concept within sociological research and describes the preference of actors in social networks to form ties on the basis of shared attributes, such as gender and race, as well as subjective characteristics such as political affiliations and desires for certain consumer goods. The study of homophily can provide important insights into the diffusion of information and behaviours within a society and has been particularly useful in understanding online community formation given the self-selected nature of the information consumed.

In this chapter we introduce the concept of homophily and show that the empirical measurement of homophily requires controlling for other factors not related to actor preferences (such as group size and endogenous network ties) that may lead to actors with shared attributes being connected to one another. We then provide a discussion of how web data can be used to advance research into political homophily, which is the phenomenon whereby people seek out others who share their political affiliation. We contend that the web provides several unique opportunities for political homophily research, but there are associated challenges that must be taken into account.

Any research involving web data for understanding social and political behaviour should first
establish that the observed online behaviour is a valid or meaningful representation of its offline counterpart – this has been referred to as the requirement that there be a ‘mapping’ between the online and offline world (Williams, 2010) or that the online data have ‘construct validity’ (Burt 2011). Our chapter discusses the construct validity of web data for political homophily and offers three tests of such validity.

One of the tests is that the web data display differential homophily, where communities exhibit idiosyncratic tie preferences within their community, rather than a uniform tendency of flocking for the population at large. In this context, we revisit the well-known ‘Divided They Blog’ 2004 weblog network data of Adamic and Glance (2005) and show how a particular statistical social network analysis technique (Exponential Random Graph Modelling) can be used to quantitatively characterise political (uniform and/or differential) homophily in the blogosphere. We use the VOSON (Virtual Observatory for the Study of Online Networks) hyperlink network research tool to construct a network of the US political blogosphere in 2011, and assess how political homophily has changed since 2004. Complementing the traditional insights gained from qualitative network visualisation techniques, we show that differential homophily has become more characteristic of the political landscape exhibited by weblogs from 2004 to 2011.

2 ASSORTATIVE MIXING AND HOMOPHILY

Assortative mixing in social networks refers to a positive correlation in the personal attributes (age, race, ethnicity, education, religion, socio-economic status, physical appearance etc.) of people who are socially connected to one another. There is strong evidence that people assortatively mix when it comes to forming friendships, marriages and sexual partnerships – this is the ‘birds of a feather flock together’ phenomenon (see, for example, McPherson, Smith-Lovin, and Cook 2001 for a review).
With regards to marriage, research reviewed in the aforementioned McPherson, Smith-Lovin, and Cook (2001) shows that Americans exhibit a preference for ‘same-race alters’ far in excess of preference for similarity based on other characteristics such as age and education. There is also evidence that people assortatively mix on the basis of political preferences, and recent work by Alford et al (2011) shows that the correlation between spouses’ political attitudes is larger than for other personality and/or physical traits.

A tendency towards politically homogeneous social interactions affects the degree of exposure to different political perspectives, and this can have an impact on, for example, the operation and effectiveness of municipal councils and civic associations (“cross-talk”, see for example Weare et al 2009). In addition, the concept of assortative mixing assists in the classification of political networks and the factions they represent, as discussed in, for example, Kydros et al (2012). And in the wake of the mass shooting tragedy in Newton, Connecticut in December 2012, network visualisations have permeated mass-media outlets to such an extent that the gun control debate in the United States is also best understood - and presented - from an assortative mixing perspective (compared with, for example, Stray [2013]).

Assortative mixing is thus an important and fundamental aspect of social networks, and consequently has received much research attention. However, assortative mixing is simply an empirical measure that describes the structure or composition of a social network (that is, which types of nodes have a higher probability of being connected) - it says nothing about the exact processes that have led to the formation of a particular social network. While it is reasonably easy to measure the level or extent of assortative mixing in a social network, it is much more difficult to discover why people are mixing on the basis of shared characteristics. We outline three main reasons why a given social network might exhibit assortative mixing.¹
First, there might be homophily – a term first coined by Lazarsfeld and Merton (1954) which refers to people forming a social tie because they prefer to be connected to someone who is similar to themselves. Homophily can in principle operate with respect to any attribute – physical characteristics such as race and gender, ‘cultural preferences’ over books and music, and political attitudes. However, when the person has choice over the attribute then it is harder to distinguish whether “birds of the feather are flocking together” (attributes are influencing friendship formation) or whether someone is becoming more like their friends (friendships are influencing attitudes and preferences).

Second, there are opportunity structures that influence social tie formation. In particular, group size is important: the smaller a particular group (for example, racial category) the more likely (all other things considered) that its group members will form social ties outside of the group (Blau, 1977). If group size is not controlled for, then there can be erroneous conclusions about the ‘homophilious’ behaviour of different groups. Independent of group size, the propinquity mechanism can also influence whether two people form a social tie – these shared ‘foci effects’ might relate to spatial proximity (for example, living in the same neighbourhood) or shared institutional environments (for example, working in the same organisation).– see, for example, Feld (1981) and Mouw and Entwisle (2006).

Finally, there are endogenous network effects, which are mechanisms that are not directly related to the attributes of individuals, but exert influence on social tie formation. First there is the process of sociality: two people might become friends simply because they are both social people and like to form lots of social ties. Second social networks tend to exhibit two properties: 1) reciprocity – if A extends the hand of friendship to B, there is good chance that B will reciprocate the friendship; 2) transitivity – the tendency for friends-of-friends to become friends (this is referred to as triadic closure). It has been argued by the proponents of balance theory (see, for example, Davis 1963) that
the social norms reciprocity and transitivity reduce the social and psychological strain that arises from unreciprocated ties and being in a situation where one’s friends are not themselves friends.

Reciprocity and transitivity can also impact on the measurement of homophily: if a particular group does have a genuine preference for forming in-group ties, then this preference will be amplified by the processes of reciprocity and transitivity. Furthermore, if there are differences between the extent of reciprocity and transitivity across different social groups (for example, one race has a cultural tendency to reciprocate friendships or introduce friends to each other) then this may obscure the cross-group comparison of homophily.

The problem for researchers studying homophily is that both opportunity structures and endogenous network effects can ‘mask’ the true level of homophily in a social network. Currarini, Jackson, and Pin (2009) demonstrate one approach for constructing measures of homophily where differences in group size are controlled for. Below, we demonstrate a statistical technique that provides estimates of homophily in a social network, where both group size and balance mechanisms are controlled for.

3 THE WEB AND POLITICAL HOMOPHILY

It was mentioned in the previous section that there is evidence people are more likely to be socially connected to other people who share a political affiliation. This section considers how research using web data is providing new insights into political homophily. First, we discuss how web data from social network sites (such as Facebook), blogs and microblogs (such as Twitter) provide several key opportunities--and sometimes challenges--for studying political homophily. Next, we examine evidence for whether political homophily exists on the web and if so, whether it has characteristics that are similar to political homophily in the offline world. Finally we consider whether web data may provide insights into how political attitudes are formed, and we also ask the
question: might the web itself contribute to political homophily?

3.1 Opportunities for studying political homophily

Web data provide several opportunities and challenges for social networks research. This section provides a summary, with particular focus on research into political homophily.

First, web data are created in a naturalistic environment and so there may be less problem of recall error and respondent burden with regard to the collection of social tie data. However, there is the additional problem that all relevant social network ties may not be observable to the researcher. If the research aims to, for example, understand the role of social networks in political preference formation using Facebook data, it may be that significant offline social ties are not represented in the data (for example, friends and family who are not on Facebook).

The naturalistic nature of web data also poses both challenges and opportunities for collecting data on the key attribute of interest: political affiliation. Focusing once again on the example of Facebook, data on the political preferences of an individual will only be available to the researcher if the Facebook user has decided to fill out the appropriate profile fields, and there may be something different about such individuals that make them less representative of the population under study (they may be more politically motivated or active than the average person in the population). Also there is a potential issue of the accuracy of the political preference data: with certain populations of study (for example, university students) there may be social pressure to display a particular political affiliation in the Facebook profile that doesn’t reflect the person’s true political preferences.

With this caveat in mind a second major advantage of web data is that it is often possible to collect complete network data (where links between all actors in the network are recorded). This allows the computation of both node-level metrics (such as degree, betweenness and closeness centrality) and
network-level metrics (such as density and centralisation) that may be important to understand the phenomenon being researched. Whole network data are necessary for being able to model ‘supra-dyadic’ phenomena, that is, where it is not just the direct ties between a person (‘ego’) and his or her social contacts (‘alters’) that are important in understanding that person’s behaviour or outcomes, but also the connections between alters themselves (and, indeed, connects between people more two or more degrees of separation from ego).

However, while Facebook might provide an opportunity to collect complete network data for a particular population, for example US college students, it needs to be recognised that this population cannot be representative of the general population. Hence, conclusions that are drawn about the extent of political homophily amongst college students may not be able to be generalised to a wider population.

It also needs to be recognised that for some web data sources, it may not be feasible to identify a bounded population from which to collect complete network data. For example, in research on the blogosphere, investigators often need to use ‘snowball’ sampling in order to build their network data because there is no sampling frame from which to randomly sample observations. Non-probability sampling techniques such as snowball sampling typically cannot be used to make inferences about population statistics – it may not be valid to make strong conclusions about the extent of political homophily in the political blogosphere, for example, when snowball sampling has been used. Further, the fact that snowball sampling may be required to construct the complete network may also make it difficult to assess the population share of, for example, conservative and liberal US political bloggers and this can have implications for the measurement of political homophily (see below).

A third and final major advantage of web data for studying political homophily is the fact that many
web datasets are longitudinal: research subjects’ political and other attributes are recorded over time, as are their social network data. This opens up the possibility for studying how political preferences and social networks co-evolve over time, allowing potential insights into the social processes underlying political preference formation.

However, there are associated challenges involved with the use of time stamped web data in the context of research into political homophily. First, it has been noted that while web environments such as Facebook provide useful data for social tie formation, they are less useful as sources of data on tie dissolution: people do not tend to ‘unfriend’ in Facebook, because the costs of maintaining a Facebook friendship are minimal. Noel and Nyham (2011) have shown that homophily in Facebook friendship retention can confound causal estimates of social influence. The implication is that one needs to be cautious when using longitudinal web data (for example, from Facebook) for researching how social ties impact on political preferences, since homophily in friendship retention (people with shared political preferences are less likely to unfriend one another) can exert upward bias on estimates of the extent to which political preferences are transmitted through social networks.

Another potential problem with time stamped social network data (both online and offline) is that people can drop out of the sample over time, and if the rate of attrition is related to the political behaviour of interest then differential rates of attrition can impact on research findings. While differential rates of attrition may not be a concern when one is studying mainstream political behaviour, it may be more of a problem if the focus of study is on radical or extreme behaviour. That is, a Facebook user who has recently started engaging in radical political behaviour might be more inclined change his or her profile to private and thus become invisible to researchers, or stop using Facebook entirely, and this will impact research into social influence and political behaviour.
3.2 Is there political homophily on the Web?

For online data to provide useful insights into offline social and political phenomena, it needs to be demonstrated that the behaviour of interest online has similar characteristics as its offline counterpart. In the context of virtual worlds, Williams (2010) refers to the need for a ‘mapping’ between the virtual and real world, and Burt (2011) argues that virtual world data need to have ‘construct validity’ in order for them to be useful in social networks research:

Do social networks in virtual worlds have the same effects observed in the real world? 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. On the other hand, if social networks in virtual worlds operate just like networks in the real world, then we can use the richer data on virtual worlds to better understand...network processes in the real world. (Burt 2011, p.5)

While Williams (2010) and Burt (2011) were referring to virtual worlds such as Second Life and massively multiplayer online role-playing games (MMORPGs), the issue of construct validity is relevant for any type of web data that are used for social and political research. In the context of using web data to study political homophily, it is therefore pertinent to ask: is there political homophily on the web, and does it have the same characteristics as political homophily offline?

Adamic and Glance (2005) demonstrated that it was possible to identify assortative mixing within political weblog networks, delineating between blogs identified as politically conservative and those identified as politically liberal. The authors constructed two datasets. The first was a single day’s snapshot of around 1500 political blogs collected by searching through several blog catalog
websites, and manually coded by political affiliation (750 liberal and 726 conservative bloggers).

On February 8 2005, the authors then extracted all hyperlinks from the front page of each blog –
there was no distinction between hyperlinks made in blogrolls (blogroll links) and those made in
posts (post citations). Figure 1.1 presents a visualisation of the 1204 non-isolate weblogs, where the
‘Divided They Blog’ phenomenon identified by Adamic and Glance (2005) is clearly displayed.²

Figure 1.1: Political Blogosphere 2004 – White = ‘Liberal’, Black = ‘Conservative’, Node Size
Indicates Degree. Data are from Adamic and Glance (2005). Visualisation by the authors.

Qualitative analysis (visualisation) of the one-day snapshot clearly showed that there was
assortative mixing on the basis of political preferences in the US political blogosphere. This was not
an unexpected finding, but it is certainly a necessary one if blog data are to have construct validity for political homophily research.

A second way of testing the construct validity of blog data for political homophily research is to establish whether there is differential homophily, that is, are the mixing patterns of the two political sub-populations different in ways that make sense? Alford et al. (2005) provide evidence for the existence of two broad political ‘phenotypes’ (observable characteristics of individuals that are determined by both genes and environmental influences): ‘absolutist’ or conservative who tend to be more suspicious of out-groups, and ‘contextualist’ or progressive who exhibit relatively tolerant attitudes towards out-groups. Drawing from this, one would expect conservatives to display greater political homophily than liberals in their linking behaviour in the blogosphere, and this is therefore another way of testing the construct validity of blog data for political homophily research.

The work by Adamic and Glance (2005) also provides some insights into the extent of differential political homophily in the US blogosphere. The authors constructed a second dataset from a subset of 40 prominent weblogs (‘A-listers’), and in contrast to the larger dataset of bloggers mentioned above, this second dataset consisted of blog posts over a two month period leading up to the 2004 US Presidential election. While the larger dataset used both blogroll links and post citations as ties between bloggers, the second A-lister dataset only used post citations. Adamic and Glance (2005) argued that since blogroll links tended to get ‘stale’, post citations were a more accurate indicator of linking behaviour, and this can be seen as a methodological response to a problem that is analogous to the ‘unfriending problem’ that Facebook researchers encounter (discussed above).

Adamic and Glance (2005) found evidence that conservative weblogs tended to cite other conservative weblogs more frequently than liberal weblogs cited other liberal weblogs. In other words, the subnetwork of liberal weblogs had more connections to conservative weblogs than vice-
versa:

Through...visualizations, we see that right-leaning blogs have a denser structure of strong connections than the left, although liberal blogs do have a few exceptionally strong reciprocated connections. (Adamic and Glance 2005, p. 40)

It should again be noted that Adamic and Glance (2005) did not provide a formal test of differential homophily, but rather used evidence from visualisations in support of this thesis. One of the challenges of formally testing for political homophily using blog data is that one would need to control for differences in group size (it was noted above that different population shares can obscure true levels of homophily) and with blog data, especially when collected using snowball sampling, it is very hard to know whether the population shares are accurate.

A final test of the construct validity of web data for political homophily research, which is related to the above two, is establishing whether the observed actors are representative of the underlying population of interest. It was mentioned above that snowball sampling may not be representative of the overall population of political bloggers. But what if the objective is to use web data to say something about the political behaviour of, for example, the voting-age population and not just the subset of the voting-age population who are also political bloggers?

It needs to be recognised that political bloggers and people engaging in politically-oriented conversations on Twitter are most certainly not representative of the voting-age population. For example, Mitchell and Hitlin (2013) compared US nationally-representative survey data with data collected from Twitter in order to gauge the public response to eight major political news events and found that the response on Twitter was much more extreme (both on the political right and left), compared with that in the national polls. Any conclusions about the extent or dynamics of political homophily based on blog or Twitter data need to be qualified, given the sampled observations are
not generally representative of the offline population.

The above discussion on the construct validity of web data for political homophily research has focused on blog data. However, political homophily research has also used other types of online data. For example, Gaines and Mondak (2009) found evidence of ideological clustering in a subset of members of the ‘UIllinois’ Facebook network, which consists of registered users affiliated with the University of Illinois at Urbana-Champaign.

Huber and Malhotra (2012) used data from a US online dating site to assess the extent of political homophily and found that “people find those with similar political beliefs more desirable and are more likely to ‘match’ with them compared to people with discordant opinions ” (p. 1). The authors argued that online dating sites provide a unique source of data for political homophily research because they allow the research to observe political preferences (expressed in profiles on the dating site) before sorting occurs. This overcomes a limitation of, for example, data on married couples where shared political affiliation may have resulted from homophily (common political preference was a factor in their union), shared environment (the couple were exposed to similar exogenous factors, for example, mass advertising or changes in the economy) or attitude conversion (one member of the couple influenced the other to change political preferences).

3.3 Social influence and political affiliation

Unlike some individual attributes (such as age and race), political preferences are not immutable: a person can change his or her political affiliation, and a potential source of change is influence from people in his or her social network. But identifying social influence using observational data is not straightforward: without detailed time stamped data on both behaviour and social networks, it is difficult to know whether two people share an attribute such as political affiliation because of social influence (one person influenced the other person to change political preferences), social selection
or homophily (the two people became socially connected because of their shared political preferences) or because both people were exposed to the same exogenous or environmental conditions that influenced them to jointly change political affiliation.

There is an active agenda of research into social influence using both offline and online social network data. With regard to online data, for example, there has been research on the spread of health behaviour in an online health community (Centola, 2010) and product adoption in an instant messaging network (Aral, Muchnik and Sundararajan, 2009). While we are not aware of a study that has attempted to understand social influence and political affiliations using web data, the offline study by Lazer et al. (2008) may provide insights into how such research could be conducted using time stamped web data. Lazer et al. (2008) collected data on research subjects’ political views before and after their exposure to one another, and they argued that this allowed them to show how social interactions influence political views.

Finally, there has also been interest in whether the extent of political homophily may be in fact influenced by the web. In the early days of the web, two radically different predictions regarding the impact of the web on politics were advanced. Some (for example, Castells 1996) argued that the web would lead to a new era of participatory democracy (broad participation in the direction and operation of the political system). In contrast, authors such as Putnam (2000) and Sunstein (2001) argued that the web would lead to increased isolation and the loss of a common political discourse, leading to cyberbalkanization – a fragmenting of the online population into narrowly-focused groups of individuals who share similar opinions and are only exposed to information that confirms their previously-held opinions. In this context, Hargittai, Gallo, and Kane (2008) used a dataset on A-list political bloggers to test whether the amount of cross-ideological linking among blogs is declining over time (this is proposed as a direct test of the ‘fragmentation’ hypothesis), and found no support for this hypothesis over a 10 month period.
4 ERGM ANALYSIS OF HOMOPHILY IN THE POLITICAL BLOGOSPHERE

While Adamic and Glance (2005) had a large impact on the search for and analysis of homophily on the web, there is to our knowledge no follow-up research to indicate 1) whether or not differential homophily is a continuing phenomenon within the political weblog community, and 2) whether or not network formation models such as ERGM can provide a good fit for such models, that is, whether or not quantitative evidence for differential homophily can be obtained.

To address both issues we revisit the political weblog phenomenon originally treated by Adamic and Glance (2005). We collect new weblog data in 2011 using a similar data collection technique, and then estimate several Exponential Random Graph Models (ERGMs) to ascertain whether differential homophily is statistically significant. In addition, we apply the same ERGM approach to the original weblog data set from Adamic and Glance (2005). We find that differential homophily is demonstrated in the updated dataset, and that the ERGM analysis indicates support for its statistical significance.

4.1 Data and Methodology

Weblog data for politically conservative and politically liberal weblogs were collected on October 27, 2011, from two websites which catalog weblogs, ‘BlogCatalog’ (www.blogcatalog.com) and ‘eTalkingHead’ (www.etalkinghead.com). These two sites provide catalogs of ‘conservative’ and ‘liberal’ websites, and provided a total of 973 unique websites, similar to the c. 1464 such websites collected by Adamic and Glance (2005).

With the explosion of weblogs over the past decade there is no guarantee that these sites would necessarily hyperlink to each other – to help circumvent this, Adamic and Glance (2005) ‘snowballed’ their original sample by an additional 30 weblogs which were cited at least 17 times by the initial collection. As the total number of weblogs collected was then 1494, these additional
30 weblogs comprise about 2% of the total. By contrast our approach does not utilise blogroll data to examine numbers of citations for snowballing the initial collection of 973 weblogs, and we take these sites as given. The premise here is that there may be a degree of self-selection by weblogs to have their data collected and stored by BlogCatalog and eTalkingHead, and that this self-selection is an *ex ante* form of snowballing at the level of the blog catalog. Future research can, however, take the 973 sites as ‘seed’ sites for a larger analysis of hyperlink citations.

The 973 weblogs so collected were then analysed by the VOSON software for hyperlink network analysis\(^4\) to generate a full network between them. When isolated weblogs are omitted, the resulting web network is depicted in Figure 1.2, with the circles depicting weblogs and the ties depicting citations (hyperlinks) to or from other weblogs. The size of the circles reflects how many of these citations there are for each weblog, or the weblog’s ‘degree’. Without the snowballing methodology of Adamic and Glance (2005), the number of weblogs which cite at least one other weblog (or which are cited by at least one other weblog) is 315, or about one-third the number of sites initially collected. The figure clearly depicts a separation between weblogs which labelled themselves as ‘conservative’ and those which labelled themselves as ‘liberal’, and reproduces the qualitative feature of Adamic and Glance (2005), Fig. 1 p. 37.
Although the graphical depiction is promising, it is unclear whether or not the separation between conservative and liberal weblogs is a symmetric or asymmetric feature, that is, the graph cannot immediately provide information about differential homophily. To answer this question requires fitting a model of network formation, and we select ERGM fitting for what follows.

4.2 Measuring Homophily: Controlling for Endogenous Network Effects

Controlling for balance mechanisms in social networks requires the use of a particular statistical social network analysis technique called Exponential Random Graph Modelling (ERGM), which is a technique for statistically ‘unpacking’ social networks. While a detailed introduction to ERGM is beyond the scope of this paper, it is useful to understand what ERGM is designed to do.
ERGM is a statistical technique that allows for the explicit modelling of the dependence among the units of observation, that is, network ties or dyads. By means of illustration, say we have three people: Ann, Sue, and David. Assume that Ann and Sue are friends and Ann and David are friends. Earlier statistical approaches to modelling social networks required the implausible assumption that the probability of Sue and David forming a friendship is the same as it would be if Ann wasn’t friends with either of them. This is implausible because it overlooks a basic mechanism in social behaviour, *triadic closure*, which is the tendency of friends of one individual to become friends themselves.

There are two types of features in social networks. Above we introduced *endogenous network effects* – these are network ties that have nothing to do with actor attributes, but are more to do with social norms. The second major feature in social networks is *actor-relation effects* – these are network ties that are created because of the characteristics or attributes of actors. There are three further sub-types of actor-relation effects:

- *Sender effects* show the impact of presence or absence of a particular actor attribute on the propensity to create, or ‘send’, ties (a significant and positive sender effect indicates that actors with the attribute send more ties than expected by chance).
- *Receiver effects* are analogous to sender effects, but refer to the propensity of receiving ties.
- *Homophily effects* occur when actors with an attribute are more likely than chance to send ties to other actors who also share the attribute.

To illustrate the above ideas, assume we are conducting an analysis of friendship formation in a school, and we have collected information on ‘friendship nominations’ (for example, person $i$ nominates person $j$ as a friend), gender and age. Assume that three of the actors are: $i$ (female, 12 years old), $j$ (female, 12 years old) and $k$ (female, 14 years old) and the friendship nominations
between these actors are: \(i\) nominates \(j\), \(j\) nominates \(k\), \(i\) nominates \(k\).

These three network ties together form a transitive triad, which is a very common structure in social networks. Why did this transitive triad form? In particular, what social process is behind the tie from \(i\) to \(k\)? There are two potential reasons. First, the tie from \(i\) to \(k\) could be an actor-relation effect (\(k\) is older than \(i\) and children are often keen to show that they hang out with the ‘big kids’; \(k\) also shares \(i\)’s gender and there is strong evidence of gender homophily especially amongst children). However, the tie from \(i\) to \(k\) could also be an endogenous network effect: the fact that \(i\) nominates \(j\) and \(j\) nominates \(k\) means there is a higher likelihood that \(i\) will also nominate \(k\) (‘a friend of my friend is also my friend’). The main point is that it is difficult to discern actor-relation effects from endogenous network effects if the social network is only observed at a single point in time. We do not know if \(i\) became friends with \(k\) first, or after \(j\) became friends with \(k\) (for example, \(i\) might have met \(k\) via \(j\)).

An exponential random graph model is essentially a pattern recognition device which breaks a given network down into all of its constituent network ‘motifs’ or ‘configurations’, and then tests whether particular configurations occur more (or less) frequently than would be expected by chance alone. In a manner similar to standard regression techniques, the model estimation produces parameter estimates and associated standard errors and confidence estimates. If a particular network motif occurs at greater or less than chance levels, we can then infer that the associated social relation has had a significant role in the development of the social network. Not controlling for endogenous or self-organising network properties may lead to a spurious conclusion that the attributes of actors are driving social tie formation when in fact it is endogenous self-organisation. In the context of homophily research, ERGM allows one to ‘control’ for endogenous network effects such as reciprocity and transitivity and thus accurately measure the homophily actor-relation effect.
4.3 ERGM Analysis

We fit five different Exponential Random Graph Models (ERGMs), in order to test the hypothesis that the political blog dataset exhibits differential homophily. The baseline model (Model 1) excludes homophily from consideration and simply assumes that the likelihood of the given network realisation is a function of the edges (‘ties’) in the system, along with the number of reciprocated (or ‘mutual’) ties. Next, we fit a model of homophily without distinguishing between weblog type (Model 2)—this is the most naïve model for homophily, as it treats both conservative and liberal weblogs symmetrically and only asks if assortative mixing is observed. Model 3 refines this by allowing for differential homophily, that is, it fits a model whereby different rates of attachment are distinguished by weblog type.

Although Model 3 may appear to be sufficient to treat the question of the existence of differential homophily in the data, there is the possibility that a model without such behaviour nevertheless appears to exhibit homophilic behaviour. This is because of the scale or group size effect mentioned above— if one community of weblog types is much larger than the other, then any modelling environment (such as ERGM) where links are formed at random will tend to link weblogs to weblogs of the larger community type. This leads to the result that the smaller community appears to have a different rate of attachment. In other words, it exhibits homophily different from the larger community, when in fact this is an artefact of random network formation.

To alleviate this requires an additional statistic to capture and control for the relative size effects of the two communities—this statistic counts the total number of times a weblog belonging to a particular community is connected, as a directed inbound link, to any other weblog. Thus, larger communities will receive more connections, and this statistic than then be used to adjust the relative degree of differential homophily. Model 4 contains both the differential homophily statistic and the
community link count (inbound) statistic.

Finally, we would like to ask how ‘dense’ these communities of weblogs are within the network, by appealing to ‘clustering’ or ‘triangle formation’. Unfortunately the MLE techniques used to estimate ERGMs can lead to a degenerate outcome, if the model is misspecified by assuming homogeneous behaviour across the network for clustering and triangle formation. Hence, we use the geometrically weighted edgewise shared partner distribution (GWESP) statistic to allow for heterogeneous triangle formation – this statistic counts the number of shared partners by edges between two weblogs, and so creates a distribution rather than a single network-wide attribute. This has been shown to alleviate the degeneracy problem in MLE over a wide range of ERGM estimation problems. We fit a model both with (Model 5a) and without (Model 5b) reciprocated links, as a priori we might expect that GWESP captures link formation better than reciprocation if cluster formation (rather than bilateral relationship formation) is more dominant in the network.

The models were fit using the ERGM routine from the ‘statnet’ (2003) package for R (2011). The results for all five models are presented in Table 1.1. From this we may infer that Model 5 (a and b), incorporating differential homophily corrected for scale and heterogeneous clustering provides the best fit to the data, as both the Akaike Information Criterion (AIC) and the log-likelihood are significant improvements over Models 1 – 4. We also see that in Model 5 (a and b) the log-odds for link formation are influenced to a greater degree by conservative weblog homophily than by liberal weblog homophily, reinforcing the notion that (on the margin) conservative weblogs reference other conservative weblogs more often than liberal weblogs reference other liberal weblogs.

We note that in Model 5a, where GWESP and reciprocating links are both offered as explanatory statistics, reciprocation ceases to be statistically significant. Moreover, Model 5b (where reciprocation is removed from the model) has the lowest AIC and highest log-likelihood of all
models, which provides evidence in favour of selecting this model as the best ERGM fit to the data.

Table 1.1: ERGMs for 2011 Political Weblog Dataset

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5a</th>
<th>Model 5b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td>−4.89453**</td>
<td>−6.4380**</td>
<td>−6.4353**</td>
<td>−6.9334**</td>
<td>−7.0349**</td>
<td>−7.0007**</td>
</tr>
<tr>
<td>Reciprocated Links</td>
<td>2.08626**</td>
<td>1.6376**</td>
<td>1.5448**</td>
<td>1.5768**</td>
<td>−0.1146</td>
<td>−</td>
</tr>
<tr>
<td>Political Spectrum</td>
<td>−</td>
<td>2.0942**</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>(Symmetric)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Spectrum, Conservative</td>
<td>−</td>
<td>−</td>
<td>1.8277**</td>
<td>2.3395**</td>
<td>2.1723**</td>
<td>2.0755**</td>
</tr>
<tr>
<td>Political Spectrum, Liberal</td>
<td>−</td>
<td>−</td>
<td>2.5745**</td>
<td>2.1817**</td>
<td>1.5105**</td>
<td>1.4624**</td>
</tr>
<tr>
<td>Indegree Count, Liberal</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>0.8824°</td>
<td>0.8537°</td>
<td>0.8132°</td>
</tr>
<tr>
<td>GWESP (τ=0.2)</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>1.7131**</td>
<td>1.7851**</td>
</tr>
<tr>
<td>AIC</td>
<td>9077.5</td>
<td>8573.8</td>
<td>8491.6</td>
<td>8482.3</td>
<td>7546.3</td>
<td>7527.9</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−4536.8</td>
<td>−4283.9</td>
<td>−4241.8</td>
<td>−4236.1</td>
<td>−3767.1</td>
<td>−3759.0</td>
</tr>
</tbody>
</table>

** < 0.1% significance; * < 1% significance

By comparison, results of the same analysis for Adamic and Glance’s (2005) data set from 2004 are presented in Table 1.2. We note again that Model 5 provides the best fit, again according to the AIC and log-likelihood – in this case, including reciprocating links was a strong log-odds predictor of link formation, and so only one version of Model 5 was fit. For the 2004 data, although there is evidence for differential homophily, it is weaker than for the 2011 data: for instance, it is roughly 8 times more likely that a conservative weblog will link to another conservative weblog in 2011, while in 2004 it is only about 1.75 times as likely. The values for liberal weblogs are about 4.3 times as likely (2011) and 2 times as likely (2004), respectively. The effect of including the GWESP statistic was roughly the same for both the 2004 and 2011 models, as was the impact of links (of any kind).

Table 1.2: ERGMs for 2004 Political Weblog Dataset
<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coefficient</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links</td>
<td>−4.6022***</td>
<td>−5.1064***</td>
<td>−5.1096***</td>
<td>−5.0771***</td>
<td>−5.7380***</td>
</tr>
<tr>
<td>Reciprocated Links</td>
<td>3.4742**</td>
<td>3.3725**</td>
<td>3.3557**</td>
<td>3.3574**</td>
<td>2.4980***</td>
</tr>
<tr>
<td>Political Spectrum</td>
<td>−</td>
<td>0.8062**</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>(Symmetric)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Spectrum,</td>
<td>−</td>
<td>−</td>
<td>0.9382**</td>
<td>0.9017**</td>
<td>0.5517**</td>
</tr>
<tr>
<td>Conservative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Spectrum,</td>
<td>−</td>
<td>−</td>
<td>0.7608**</td>
<td>0.7952**</td>
<td>0.7168**</td>
</tr>
<tr>
<td>Liberal</td>
<td></td>
<td></td>
<td>−</td>
<td>−0.0588+</td>
<td>−0.0517**</td>
</tr>
<tr>
<td>Indegree Count,</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Liberal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWESP (τ=0.2)</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>1.2440**</td>
</tr>
<tr>
<td>AIC</td>
<td>191646</td>
<td>188372</td>
<td>188215</td>
<td>188233</td>
<td>166127</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−95820.9</td>
<td>−94182.9</td>
<td>−94103.6</td>
<td>−94111.7</td>
<td>−83057.3</td>
</tr>
</tbody>
</table>

** < 0.1% significance; * < 1% significance; + < 10% significance

5 CONCLUSION

This chapter has focused on how homophily, as a major contributor to assortative mixing on the basis of political affiliation, may be conceptually defined and empirically identified in web data. It is worth noting that we have used as our point of departure the existence (or lack thereof) of the characteristic of homophily, rather than first specifying the micro-level choices of individuals comprising a web network that would lead to homophily. By so doing, we have replaced the actual decision-making process of individuals, which is exceptionally complex for networks of the size and variety discussed here, by a behavioural assumption – that ‘like prefers like’. But this assumption nevertheless carries with it a rich set of implications for network types and topologies, which can replicate many sets of interactions observed in real networks.

This chapter first provided a background on homophily and its measurement, showing how homophily (people having a preference to be connected to other similar people) differs from assortative mixing (an empirical statement of whether like are connected to like), and is in fact one of several reasons why a given social network may display assortative mixing. We then focused on how web data can provide new insights into political homophily and discussed the opportunities (and challenges) of web data in this context. The concept of ‘construct validity’ of web data was
introduced and we provided three tests of whether web data have construct validity for political homophily research. We also briefly discussed how web data are being used for research into social influence, and pointed to future opportunities for such research in the context of political affiliations.

As part of this review, we also examine the role that digital media may play in increasing the fragmentation of views within a population, by allowing those of similar opinions to more readily ‘flock together’, a process which has become known as ‘cyberbalkanization’. The resulting emergence of narrow online communities of like-minded interests is a growing concern for policy-makers due to the perceived potential of weakening levels of social cohesion, inclusion and tolerance.

Our review of homophily and its measurement identified the need for statistical social network analysis for conducting rigorous test of homophily. We demonstrated how the ERGM technique can be used for quantifying homophily in the context of political weblogs, and applied ERGM to both the weblog data collected by Adamic and Glance (2005) and a more current dataset that was constructed for this chapter. Provided that the fidelity of the data is high enough to provide meaningful statistical inference, we contend that the web will increasingly be a source of data for political homophily research, and we have shown that web collection and analysis tools such as those provided by VOSON are of great importance in obtaining this quality of data.
Robert Ackland is an Associate Professor in the Australian Demographic and Social Research Institute at the Australian National University. He has degrees in economics from the University of Melbourne, Yale University (where he was a Fulbright Scholar) and the ANU, where he gained his PhD in 2001. While Robert’s earlier career was in applied economics, since 2002 he has been working in the fields of network science, computational social science and web science, with a particular focus on quantitative analysis of online social and organisational networks. His research has appeared in journals such as the Review of Economics and Statistics, Social Networks, Computational Economics, Social Science Computer Review, and the Journal of Social Structure. Robert leads the Virtual Observatory for the Study of Online Networks project (http://voson.anu.edu.au) and teaches on the social science of the Internet, statistics and online research methods. He has been chief investigator on four Australian Research Council grants and in 2007, he was a UK National Centre for e-Social Science Visiting Fellow and James Martin Visiting Fellow at the Oxford Internet Institute. His book Web Social Science (SAGE Publications) will be in bookstores in July 2013.

Jamsheed Shorish is Chief Technology Officer at Uberlink Corporation and founder and CEO of Shorish Research. He is a computational economist, earning his B.A. from Carleton College (MN, USA) and M.Sci. and Ph.D. from Carnegie Mellon University (PA, USA). Prior to his entrepreneurial activities he was Science Officer for the Information and Communication Technologies Domain at COST (the European Cooperation in Science and Technology) in Brussels, and a professor at the Institute for Advanced Studies, Vienna and the University of Aarhus, Denmark. Jamsheed has a background in complex systems analysis and dynamic optimisation, with published work appearing in journals such as the Journal of Economic Theory, the Journal of Economic Dynamics and Control, Economic Theory and Computational Economics. He is currently
an adjunct faculty member with the Australian Demographic and Social Research Institute at the
Australian National University and an associated faculty member at the Institute for Advanced
Studies, where he teaches on a regular basis. In 2007 he was a UK National Centre for e-Social
Science (NceSS) Visiting Fellow at the University of Manchester.
REFERENCES


1 See Wimmer and Lewis (2010) for a detailed review.

2 Visualizations in this chapter were created using Gephi (Bastian, Heymann, and Jacomy 2009), an open-source network visualization platform.

3 Ackland and Shorish (2009) propose an economic model of link formation that can explain the existence of differential political homophily in the blogosphere, where the link formation behavior of bloggers is influenced by the underlying distribution of political preferences.

4 For more information on the VOSON software and associated methods see, for example, Ackland (2011), Ackland and O’Neil (2011) and Ackland and Gibson (2013).


6 See Lusher and Ackland (2011) for an introduction to using ERGMs to study hyperlink networks.

7 R v. 2.13.2 64-bit on OSX 10.7.2. MLE and MCMC with sample size of 1 million per iteration, max. 15 iterations for each model.

8 Data Source: Adamic and Glance (2005)

9 Compared with Ackland and Shorish (2009), where such micro-level foundations are explored.