Network Formation in the Political Blogosphere: An Application of Agent Based Simulation and e-Research Tools

Robert Ackland
Jamsheed Shorish

IRIHS INSTITUTIONAL REPOSITORY AT IHS
Vienna

INSTITUT FÜR HÖHERE STUDIEN
INSTITUTE FOR ADVANCED STUDIES
Vienna

Network Formation in the Political Blogosphere:

An Application of Agent Based Simulation and e-Research Tools

Robert Ackland, Jamsheed Shorish

Network Formation in the Political Blogosphere:

An Application of Agent Based Simulation and e-Research Tools

Robert Ackland, Jamsheed Shorish

October 2007

Contact:

Robert Ackland
Australian Demographic and Social Research Institute,
The Australian National University
Canberra ACT 0200, Australia
email: Robert.Ackland@anu.edu.au

Jamsheed Shorish
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
email: shorish@ihs.ac.at
and
University of Illinois at Urbana-Champaign, USA

Abstract

The political blogosphere has recently been the focus of attention for social network analysis and applications of network and graph theory. In a recent paper, Adamic and Glance (2005) report differences between the linking behavior of politically conservative vs. politically liberal Web bloggers. We construct a simple agent-based network formation model which shows that one such difference, demonstrating what we term 'political homophily', can be generated by connecting the blogosphere to the underlying population distribution of political preferences. The model is implemented as a web service in the e-tool VOSON (Virtual Observatory for the Study of Online Networks), and both model and tool serve to define a natural environment for research into link formation behavior with large numbers of heterogeneous network participants.

Keywords

Network formation, social network analysis, blogosphere, VOSON, agentbased simulation

JEL Classification

D85, C63, L86

Contents

1 Introduction

In graph theory, "assortative mixing" is the extent to which nodes connect preferentially to other nodes with similar characteristics – see, for example, Newman (2002). A particular version of assortative mixing is where new entrants to a growing network prefer to connect with existing nodes who have higher levels of connectivity (e.g. indegree or outdegree), and Barabási and Albert (1999) have promulgated an influential model of "preferential attachment" which can be used to explain the existence of power laws in the distribution of links or edges in large-scale socially-generated networks such as the WWW.

A network is also said to exhibit assortative mixing when link decisions are based on non-graph theoretic node attributes (in the sociology literature, this is known as homophily); in sexual networks, for example, individuals tend to choose partners who have similar age, race and socioeconomic status. Political preference is another individual-level characteristic that we expect may influence networking behavior, and the emergence of the political 'blogosphere', a community of political commentators who use chronologically updated web pages ('weblogs' or 'blogs') to provide commentary, provides a unique data source for studying politically-oriented assortative behavior.

A key feature of blogging is web page linking ('hyperlinking') to blogs from other commentators and to web pages from traditional news sources (newspapers, wire services, etc.). Adamic and Glance (2005), in their analysis of political blogger activity during the 2004 U.S. presidential election, reported differences in the hyperlinking behavior of bloggers classified as politically conservative, compared with their politically liberal counterparts. Specifically, conservative bloggers were found to form a denser network of strong connections, and while Adamic and Glance (2005) did not directly test for differences in assortative mixing by conservatives and liberals, their results are suggestive of conservatives being relatively more inclined to link to their own in-group.[1]

---

[1] In a study of the web presence of U.S. participants in the abortion debate, Adamic (1999) found that the network of pro-life web sites is more "tightly-knit" compared with the pro-choice network; this provides additional evidence that conservatives and liberals network differently, but more empirical analysis is warranted.

If it *is* the case that conservatives exhibit a higher degree of "political homophily" on the web, then what could account for this behavior? Alford, Funk, and Hibbing (2005) provide evidence for the existence of two broad political phenotypes: "absolutist" or conservative who exhibit a suspicion of out-groups and seek in-group unity, and "contextualist" or progressive who exhibit relatively tolerant attitudes towards out-groups. The existence of a conservative "type" who is genetically pre-disposed to exhibiting a greater degree of homophily could thus be used as an explanation for differential assortative mixing in the political blogosphere.

In the present paper, we provide an alternative explanation for differences in political homophily in the blogosphere via a simple network formation model which maps the underlying population distribution of political types to the linking behavior among and between bloggers. We show that if a blogger's political preference is in the minority of the population as a whole, then they will exhibit a higher degree of political homophily, leading to the formation of smaller, more homogenous networks. By contrast, those with the majority preference will form more attachments to bloggers from the other side of the political spectrum, leading to larger and more heterogeneous networks.

We develop the model along the same lines as those surveyed in Jackson (2004) and explored more fully in e.g. Jackson and Rogers (2005), in which individual agents (here, bloggers) are faced with both benefits and costs of link formation. Conditional upon these benefits and costs, and upon the information at their disposal, bloggers make optimal decisions about whom to link to (rather than having their attachment behavior 'hard-coded' as in Barabási and Albert 1999). Overall, it is the interaction between these link valuations and the *incomplete* information at the disposal of the blogger which generates the higher degree of assortative mixing exhibited by the minority political preference.

Our research agenda currently comprises three strands. First, we are conducting a thorough empirical analysis (drawing primarily from social network analysis, SNA) of the evolution of the U.S. political blogosphere. The work of Adamic and Glance (2005) is informative, but is limited to A-list blogging activity during the 2004 U.S. presidential election, while we are

2

interested in studying the activities of A-listers, B-listers and new entrants over a longer period. Second, in order to validate our model we will need estimates of the distribution of political preferences of the online population at given points in time. These two strands are work in progress.

The third strand of our work is the development of the model itself, and this is the focus of the present paper. As the model is dynamic, contains many heterogeneous agents, and is difficult to analyze in full detail analytically (but see Ackland and Shorish 2007 for an equilibrium treatment with incomplete information), we use simulations and SNA tools to understand the connection between the underlying population distribution of political preferences and the resulting blogosphere structure. Our agent-based model is written in Python, and in order to access existing SNA tools to analyze the simulated link data for various parameterizations of the model, we implement the model as a web service that interacts seamlessly with SNA and mapping routines available within the VOSON (Virtual Observatory for the Study of Online Networks) e-Research tool (Ackland, O'Neil, Standish, and Buchhorn 2006). This strategy of running the Python model as a web service accessible from VOSON also fits with our long-term goal of accessing high performance computing (HPC) resources to enable us to run larger-scale and more complex models.

In summary, we present our research as an innovative example of an economic analysis of online networking behavior, where the research methodology itself is benefiting from emerging e-Research/cyberinfrastructure technologies.

2   The Network

2.1   Agents and Agent Preferences

An agent new to the 'blogosphere' network is drawn at random and connects to a subset of other nodes in the network, in the following way. If the agent is of type J ('left' or 'right'), the agent prefers to be attached with another agent of the same type more than an agent of a differing

3

type. This 'preferential attachment' is an abstraction of the notion that a new blogger wishes (ultimately) to be recognized by the blogging community of like type, and thus links to blogs of like type in the hope that reciprocal attachment will obtain.

We assume an extreme form of preferential attachment: if an agent forms a link to another agent of their own type, they derive a benefit $k > 0$ from the attachment. Otherwise, they derive no benefit.

In addition to the 'cost' (no benefit) of linking to an incorrect type, there is also a direct cost of forming a link regardless of type. If on the contrary links were costless, it would be optimal for the agent to form as many links as there are agents of a given type, in order to maximize the chance that a link is reciprocated. In reality this behavior is not observed, indicating that there exists a cost to link formation prohibiting excessive linking from taking place. Intuitively this is probably a consequence of the linear nature of blog link lists–since blog links are not randomly retrieved by readers, but are instead processed in a linear fashion as rendered by the browser, there is a cost borne by the reader of having to wade through a large list of links. Knowing this, most bloggers limit the number of links they provide to something far less than the number of links which could technically be provided.

We generalize this cost of linking (as is common in the literature) as a cost of $c > 0$ per link which is borne by the linker. In our parlance, the cost is equal to the 'strength' of the link formed, i.e. $c = f(i, j)$ for a link from i to j–this allows a degree of freedom to the linker, in that links can be relatively low cost or high cost, similar to menu costs.

Lastly, we assume that agents are risk-neutral, so that the expected payoff of forming a link may be given by:

$$U := P(J) * (k - c) + (1 - P(J)) * (-c) = kP(J) - c, \qquad (2.1)$$

where $P(J)$ is the probability that a link is formed to an agent of the same type.

2.2 Kinds of Agents

We start with three *kinds* of agents, each of which may be one of the two aforementioned *types*. The first agent is a new **entrant** into an existing network–the entrant has a type t which is known to the entrant but unknown to the other two kinds of agents. The second agent is an existing member of the network, known as a **B-list** member. The B-list member prefers to link to agents of the same type. Lastly, there is another existing network member called an **A-list** member. The A-list member is similar to the B-list member in that it prefers to link to an agent of the same type, but has different preferences.

2.3 Entrant Behavior

The entrant is assumed to create an out-link to an A-list member of their type. The entrant wishes to have links reciprocated, and a fully dynamic model would have link choices dependent upon the probability of receiving a link in the future. Also in a dynamic model (see below), an entrant becomes a B-list member in the period following entry.

2.4 A-list and B-list Behavior

The A-list and B-list members derive utility from linking to entrants of their own type. If their type is given by J, the payoff from forming a link for an A-list member is

$$U_a = k_a \mathrm{P}(t = J | \mathscr{F}_a) - c_a, \tag{2.2}$$

where k_a and c_a are scalars ('benefit' and 'cost') such that $k_a > c_a$, and $\mathrm{P}(t = J | \sigma)$ is the probability that (conditional upon the information set $\mathscr{F}_a$ available to the A-list member) the type t of the entrant matches the A-list's type J.

Similarly, a B-list member's link to an entrant carries the payoff

$$U_b = k_b \mathrm{P}(t = J | \mathscr{F}_b) - c_b. \tag{2.3}$$

5

An A-list or B-list member will form a link if and only if

$$U_a > 0, U_b > 0.$$

What distinguishes A-list and B-list members are (1) their respective scalars $k_{a,b}$ and $c_{a,b}$, and (2) the information sets $\mathscr{F}_{a,b}$. In what follows we shall see that much of our hypothesis can be confirmed by concentrating on (1), and letting the information sets of A-list and B-list members be the same (this can be relaxed in more complicated models).

As the backdrop for this analysis is the distinction between A-list and B-list bloggers, we assume that these two kinds of bloggers are distinguished by their opportunity cost of forming a link to another agent, and by their respective utility benefits from such links. For example, we might suppose that as an A-list blogger has far more influence than a B-list blogger in the blogosphere, an A-list blogger will derive less (probability-weighted) utility from forming a link to a new entrant than a B-list blogger, who might value such out-links more. At the same time, the A-list blogger's blog 'real estate' is valued more than a B-list blogger's blog space, so the cost for forming a link for an A-list blogger (and hence reducing the amount of usable blog space) is higher than for a B-list member.

This intuition can be encapsulated by the following assumptions:

Assumption 2.1. *The marginal (probability weighted) benefits of forming a link for A-list and B-list members are ranked such that $k_b > k_a$.*

Assumption 2.2. *The marginal cost of forming a link for A-list and B-list members are ranked such that $c_a > c_b$.*

These assumption are summarized by the inequality:

$$r_a > r_b, \tag{2.4}$$

where $r_{a,b}$ is the cost-benefit ratio defined earlier.

2.5 Equilibrium Decisions

It is clear that when A-list and B-list members have the same information set $\mathscr{F}_a = \mathscr{F}_b = \mathscr{F}$, they will form links to a new entrant if and only if

$$P(t = J|\mathscr{F}) \geq r_s, \tag{2.5}$$

for $s \in \{a, b\}$, respectively.

3 Network Evolution Under Uncertainty: Model and Simulations

In an accompanying paper (Ackland and Shorish 2007) we discuss the analytical solution to this modeling environment, when the underlying population type distribution $P(J)$ is known. We show that in equilibrium the minority population, i.e. the type j where $P(j) < 0.5$, is more likely to form attachments to the same type rather than to the other type, leading to smaller and more homogenous networks, compared with the majority population.

If there is uncertainty over the underlying population type distribution $P(J)$ then the aforementioned equilibrium outcomes will in general not obtain, as agents will make decisions based upon their 'best guess', i.e. the model is one of imperfect (but homogeneous) information. In this case it is the interplay between the random type distribution and the heterogeneous cost-benefit structure between A- and B-list agents which determines the overall network structure. As before, the networks which emerge serve to underscore the empirical conclusion that minority groups tend to form smaller, more densely connected networks.

To simulate the evolution of the blogosphere from an initial underlying population we adopt an agent-based model, in which 'agents' are individual bloggers categorized by kind, as "A" or "B" listers. These bloggers have imperfect information regarding the population distribution of types (e.g. Republican vs. Democrat), which is assumed to evolve stochastically over time as

a mean-revering process. Knowing their own cost-benefit ratio, bloggers decide whether or not to link to a new blogger (an 'entrant') according to whether or not their expected benefit from doing so exceeds their expected cost.

In the simulations time $t = 0, 1, \ldots$ is discrete. Every agent can be one of two types ("Zero" and "One"), and may be of three different kinds ("A"-lister, "B"-lister, or "Entrant"). Each period a single agent is drawn from the population probability density P_t, where P_t = probability that the agent drawn is of type Zero (and hence $1 - P_t$ is the probability the agent is of type One).

The evolution of P_t over time, or the population distribution's generating process, is assumed for simplicity to be an AR(1) stochastic process with uniformly distributed shocks:

$$P_{t+1} = \delta P_t + \varepsilon_{t+1}, \tag{3.1}$$

where $\delta \in (0, 1)$ is the mean-reversion parameter and ε is the random shock.

In order to keep $P_t \in [0, 1] \; \forall t$, i.e. to keep P_t a probability, certain restrictions must be placed on the variance of the distribution for ε. In addition, it is convenient to parameterize the population distribution by its long-run value, denoted $\bar{P}$, which is the unconditional expectation of the AR(1) process over time. This parameterization means that in the long run,

$$\bar{P} = \frac{1}{1 - \delta} \mathbb{E}(\varepsilon), \tag{3.2}$$

which is a restriction on ε's mean as well.

These restrictions translate into conditions upon the endpoints of the uniform probability distribution of ε, $\underline{\varepsilon}$ and $\bar{\varepsilon}$, with $\bar{\varepsilon} > \underline{\varepsilon}$. For example, if $\bar{P} < 0.5$, then

- $\underline{\varepsilon} = 0$,

- $\bar{\varepsilon} = 2\bar{P}(1 - \delta)$.

On the other hand, if $\bar{P} \geq 0.5$, then

- $\underline{\varepsilon} = (2\bar{P} - 1)(1 - \delta)$,

8

- $\bar{\varepsilon} = 1 - \delta$.

When these endpoints are used, for any parameterization of the system given by $\bar{P}$, the long-run value for the population distribution, the AR(1) process for P_t is well-defined for all periods t.

Agents in the network do not know P_t, the current population distribution–rather, they have information only up to the previous period, P_{t-1}, and they must use this information to form expectations on the likelihood that the newly-drawn entrant is of their type. Equivalently, we assume that agents know the population distribution generating process, but cannot observe ε_t until period $t+1$.

Consider the behavior of a minority agent, i.e. an agent who's type distribution is given by $P_t < 0.5$ (the analysis also goes through for $P_t = 0.5$). An agent with a cost-benefit ratio of r, then, decides whether or not to form a link according to the following probabilities:

3.1   Case 1: $\delta P_t > r$

Due to the restrictions on the distribution of ε the autoregressive shock cannot take negative values–hence $P_{t+1} \geq P_t$ and it is *always* worth forming a link.

3.2   Case 2: $\delta P_t \leq r$

3.2.1   Case 2a: $r - \delta P_t > \bar{\varepsilon}$

If $r - \delta P$ is positive, then there might exist draws of ε which 'push' δP_t over r–but if $r - \delta P_t \geq \bar{\varepsilon}$ then no such draw exists. Thus, $P_{t+1} < r$ for sure and a link is *never* formed.

3.2.2   Case 2b: $r - \delta P_t \leq \bar{\varepsilon}$

This is the most interesting case, because it can lead to stochastic behavior on the part of the agent. Here, there is an open set of values for ε which could 'push' δP_t over r. As the distribu-

9

tion for ε is uniform, this open set is actually a probability mass. In particular:

$$\text{with probability } \frac{1}{\bar{\varepsilon} - \underline{\varepsilon}}(r - \delta P_t - \underline{\varepsilon}), \ P_{t+1} < r,$$

$$\text{with probability} \frac{1}{\bar{\varepsilon} - \underline{\varepsilon}}(\bar{\varepsilon} - (r - \delta P_t)), \ P_{t+1} > r.$$

(The event $P_{t+1} = r$ has zero mass since draws of ε are from a continuum.)

In Case 2b, then, the agent performs another expected utility maximization. Conditional upon $r - \delta P_t \leq \bar{\varepsilon}$, case 2b says that the expected payoff of forming a link is (c.f. equation 2.1 above)

$$U = k\frac{1}{\bar{\varepsilon} - \underline{\varepsilon}}(\bar{\varepsilon} - (r - \delta P_t)) - c, \tag{3.3}$$

In other words, a link has positive expected value when

$$\frac{1}{\bar{\varepsilon} - \underline{\varepsilon}}(\bar{\varepsilon} - (r - \delta P_t)) > \frac{k}{c} = r \Leftrightarrow P_t > \frac{1}{\delta}(r + (\bar{\varepsilon} - \underline{\varepsilon})r - \bar{\varepsilon}).$$

Thus, when expected utility maximizing, the agent should form a link whenever

$$P_t > \frac{1}{\delta}(r + (\bar{\varepsilon} - \underline{\varepsilon})r - \bar{\varepsilon}), \tag{3.4}$$

and not form a link otherwise.

The problem for an agent of the majority type, i.e $Q_{t+1} := 1 - P_{t+1} > 0.5$ is almost identical, the only difference being that if the probability distribution of the minority type, P_t, follows the process

$$P_{t+1} = \delta P_t + \varepsilon_{t+1},$$

then the process of the majority type Q_t is

$$Q_{t+1} = \delta Q_t + (1 - \delta) - \varepsilon_{t+1}$$

and the above inequalities must be adjusted for the new shock distribution (this is a straightforward exercise, and is omitted for brevity).

4   Numerical Results

We simulate a network with $N = 102$ agents, or nodes, as discussed in Section 3 above (the extra two nodes represent initial A-list bloggers, one for each type, who initialize the network). The long-run probability distribution of types was set first at 0.5, i.e. equal measures of both types ("50-50" below), and then to 0.4, where the minority type represents 40% of the underlying population ("40-70" below). The parameters for cost and benefit were set at $c_a = 5$ and $k_a = 10$ for A-list agents, and $c_b = 4.9$ and $k_b = 10.1$ for the B-list agents. These parameterizations satisfy assumptions 2.1 and 2.2 given earlier.

As mentioned earlier, the agent-based simulation was written in Python and was then exposed as a web service (the Perl SOAP::Lite module was used to run the web service, but in the future we will investigate Python web service libraries) and accessed from the VOSON e-Research software (`http://voson.anu.edu.au`). VOSON is web-based software incorporating web mining, data visualization, and more traditional empirical social science methods. VOSON is built on a web services framework, which enables access and sharing of distributed resources such as datasets, methods and computational cycles. The main advantage of running the Python simulation as a VOSON web service is that we can then make use of SNA and network mapping routines that are also accessible as web services from VOSON. The VOSON prototype SNA service currently provides several key network measures, and is based on the sna package which is available in the R Project for Statistical Computing (a future plan is to use the R sna package for all SNA routines within VOSON).

The results reported here and in Tables I and II are indicative of the run outputs over many repetitions–future research using HPC resources will allow for greater numbers of agents to be simulated and analyzed. (In the tables, n refers to the number of nodes of each type, n_i to the

total number of inlinks, $\bar{n}_i$ to the average number of inlinks per node, n_o to the total number of outlinks, $\bar{n}_o$ to the average number of outlinks per node, $n_o(+)$ to the number of outlinks to the same type, $n_o(-)$ to the number of outlinks to the other type, and finally $\frac{n_o(+)}{n_o}$ to the fraction of outlinks which were formed to the same type. This last measure corresponds to 'political homophily' as described in the text.)

4.1   50-50 Population Results (Table I)

As expected, simulations run with both types of agent equally represented in the underlying population were symmetric. There were 48 type zero nodes and 54 type one nodes, reflecting the underlying population distribution. Both networks of each type formed links with their own type and the other type with equal number and frequency (for example, type zero agents had an average out-degree $\bar{n}_o$ of 29.5, as opposed to type one agents with average out-degree 28.01.). In addition, the frequency of linking to one's own type was essentially identical across types (54% for type zero agents, and 55% for type one agents). As discussed above, this is due to the fact that the interaction between the forecast of P_{t+1} with the cost-benefit ratios r_a and r_b was the same for both types. As both types of agents are symmetric in this case, there is no difference between them in the simulations.

Table I: 50% - 50% Population Division, $N = 102$

Type	n	n_i	$\bar{n}_i$	n_o	$\bar{n}_o$	$n_o(+)$	$n_o(-)$	$\frac{n_o(+)}{n_o}$
Zero	48	1445	30.10	1416	29.50	765	651	0.54
One	54	1482	27.44	1513	28.01	832	681	0.55

4.2   40-60 Population Results (Table II)

When there is a minority in the population, the results change substantially. There were 43 type zero agents and 59 type one agents, reflecting the underlying population distribution. In addition, the network formed by the minority population is smaller, especially with regard to

outlinks n_o: there are 220 total outlinks for type zero agents, as opposed to 3321 outlinks for type one agents. The average out-degree per agent $\bar{n}_o$ is just as striking: for type zero agents there are just 5.12 average outlinks, while average outdegree for type one agents is 56.29, a ten-fold increase. This is due to the fact that, as it is less likely a random entrant is of their type, an existing minority agent in the network will be more cautious in attaching a link to an entrant than their majority agent counterpart, and will on average form less links. Lastly, when a minority agent does form a link, it is more likely to be a link to one's own type, demonstrating political homophily. For type zero agents, 81% of all outlinks are to their own type. For the majority type one agents, however, this number falls to 53%, which is about the same as the linking frequency for a 50-50 division of the underlying population.

Again, political homophily is due to the cost-benefit structure of the model–for the majority agent, it is very likely that an entrant drawn is of their type. The expected payoff increases, then, and more links are formed, which *ex post* can include agents of the other type. As the minority agents are more circumspect, they connect more frequently to their own type, and connect to less agents in general (cf. equation 3.4 above). It is important to note here that there is no *inherent* distinction between majority and minority agents. Both types behave in precisely the same way, as given in Section 3. The difference is that they evaluate their expected payoffs differently, according to whether or not their own type distribution places them in the majority or in the minority of the population. This defines the likelihood of a new entrant being of their own type, and hence changes the likelihood of their forming a link.

Table II: 40% - 60% Population Division, $N = 102$

Type	n	n_i	$\bar{n}_i$	n_o	$\bar{n}_o$	$n_o(+)$	$n_o(-)$	$\frac{n_o(+)}{n_o}$
Zero	43	1736	40.37	220	5.12	179	41	0.81
One	59	1803	30.56	3321	56.29	1763	1558	0.53

References

ACKLAND, R., M. O'NEIL, R. STANDISH, AND M. BUCHHORN (2006): "VOSON: A Web Services Approach for Facilitating Research into Online Networks," in *Proceedings of the Second International Conference on e-Social Science*. University of Manchester.

ADAMIC, L. (1999): "The small world web," in *Proceedings of the 3rd European Conf. on Digital Libraries, volume 1696 of Lecture notes in Computer Science*, pp. 443–452.

ADAMIC, L., AND N. GLANCE (2005): "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," in *LinkKDD '05: Proceedings of the 3rd International workshop on Link discovery*. ACM Press, New York, USA.

ALFORD, R., C. FUNK, AND J. HIBBING (2005): "Are Political Orientations Genetically Transmitted?," *American Political Science Review*, 99, 153–167.

BARABÁSI, A.-L., AND R. ALBERT (1999): "Emergence of scaling in random networks," *Science*, 286, 509–512.

JACKSON, M. O. (2004): "A Survey of Models of Network Formation: Stability and Efficiency," in *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. by G. Demange, and M. Wooders. Cambridge University Press, Cambridge UK.

JACKSON, M. O., AND B. W. ROGERS (2005): "The Economics of Small Worlds," *Journal of the European Economic Association*, 3, 617–627.

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