Multiple Relationship Types in Online Communities and Social Networks

Tad Hogg and **Dennis M. Wilkinson** and **Gabor Szabo** and **Michael J. Brzozowski**
HP Labs
Palo Alto, CA

Abstract

Online social networking is increasingly popular and is a feature of many websites. Providing multiple types of relationship, such as friend, fan or colleague, can enhance the significance of the networks. We present an empirical study of an online political forum where users engage in content creation, voting, and discussion. The users also make explicit connections via three relationship types, forming three distinct networks. We establish a strong correlation between network participation and site activity and show that users stayed faithful to the relationship semantics, in aggregate. Moreover, we demonstrate significant structural differences among the networks, indicating different uses for each. Social networking features may help spread the word about new content, but we show that the networks played a surprisingly moderate role in this respect. Usage and social networking patterns were typical of many web communities, suggesting that multiple relationship types could be successfully featured in other such communities.

Introduction

Online services can provide a wide range of options to their users, particularly when aggregating content and services from many providers. This range allows catering to many specialized interests, but can make it difficult for users to find relevant content and trust the provider's quality assurances. A common approach to this difficulty provides a mechanism that builds a network among users, which is used to suggest suitable choices and providers (Goldberg *et al.* 1992; Resnick & Varian 1997; Lam 2004; Hogg & Adamic 2004). Such networks can be created implicitly from user behavior (*e.g.*, linking to other users who made similar choices to you in the past) or explicitly (*e.g.*, by users specifying others as friends).

Most online social networks allow only one type of connection, a "friendship". As the central application of a number of extremely popular services such as MySpace and Facebook, social networks have been extensively studied (Adamic, Buyukkokten, & Adar 2003; Ahn & others 2007) and shown to possess a relatively consistent set of structural characteristics (Newman 2003). Recently, online providers have included social networking components in other services, and the role of social networks in information filtering (Lerman 2007) and viral marketing (Leskovec, Adamic, & Huberman 2007) have been studied on large websites such as Digg and Amazon.

By treating all relationships in the same way, simple social networks ignore nuances of differing strengths and types of relationship among people (boyd 2006). For online services, such as recommendation (Gladwell 2000) and search (Watts, Dodds, & Newman 2002), relationship type could be significant, so social network applications may benefit from encouraging users to specify multiple relationship types. On the other hand, eliciting specific relationship information requires more user effort, which may inhibit their creating the networks. Indeed, while Facebook allows users to annotate how they met a friend, this step is often skipped.

Effective social networking using multiple relationship types requires suitable tradeoffs between the benefits of having the extra relationship information and ease of use. Evaluating these tradeoffs raises several questions:

- Do users create links of each available type? If so, do these links match their nominal semantics?
- Do the networks of different link types benefit the users by enabling them to use the services effectively or find other similar users?
- Do these networks improve performance for the service provider compared to using a single link type?

More generally, for a particular service, which types of relationship between people are most useful and how can they be elicited or inferred?

In this paper, we focus on the first two of these questions by presenting an empirical examination of online activity by all 15,419 users of a politically-oriented web community, Essembly[1], between its inception in August 2005 and December 2006. This activity includes participating in three distinct interpersonal networks as well as creating and voting on content. The three networks were a social network, an ideological preference network, and an anti-preference network, called *friends*, *allies* and *nemeses*, respectively. All links in these networks are explicitly created by users.

In the remainder of this paper, we describe our data set, followed by usage patterns, network properties, and the net-

[1] Essembly LLC at www.essembly.com

works' relation to user behavior. We conclude with a discussion of our results and possible future work.

Essembly

Essembly is an online service helping users to engage in political discussion and organization. Essembly allows members to post *resolves* reflecting controversial issues, *e.g.* "Overall, free trade is good for American workers." Other members can then vote on these resolves, using a four-point scale: *Agree*, *Lean Agree*, *Lean Against* and *Against*. Users can only vote once, and all their votes are viewable by other members, forming an ideological profile.

During their sign up, new users are encouraged to vote on ten initial resolves to create an ideological profile. Subsequent votes on other, user-generated resolves update their profiles. Votes on the ten initial resolves do not reflect any mechanism for content discovery so we do not consider them in our study of network use. For the remainder of this paper, we use "resolve" to mean the 25953 user-created resolves, and consider only votes on user-created resolves, unless otherwise noted.

The distinct social and ideological networks enable users to distinguish between people they know personally and people encountered on the site with whom they tend to agree or disagree. Network links are formed by invitation only and each must be approved by the invitee. Thus all three networks in Essembly are explicitly created by users and are symmetric (i.e., undirected). Essembly users can identify potential allies or nemeses by comparing ideological profiles. The Essembly user interface presents several options for users to discover new resolves, based on votes by network neighbors, recency, overall popularity, and degree of controversy. Essembly is well-suited to a study of multiple relationship types because it provides the networks and easily quantifiable behavior using those networks.

Our data set consists of fully anonymized voting records and network connections, representing a complete history of voting activity on the site during the study period and the final state of the three networks.[2] The data set does not include the content of the resolves, the history of link formation in the networks nor which of the two users involved initiated the formation of each link.

User behavior and network structure

In this section we present results on patterns of user activity and details of network structure. We also show the relation between user-user homophily and distance in the network.

Usage patterns

The first question we address is whether user behavior in Essembly is similar to that seen in many other social network web sites. 12,440 users voted on resolves, casting a total of 1.3 million votes. The distribution of votes per user is heavy tailed, with a few very active users making a disproportionately large number of votes: the top 5% of users cast 81% of the votes and the top 1% cast 44%. Such a distribution also occurs in other online services, notably Wikipedia editing (Wilkinson & Huberman 2007), Digg recommendations (Lerman 2007), and Facebook messaging (Golder, Wilkinson, & Huberman 2007).

Our data has 24,953 resolves, averaging 52.6 votes. Fig. 1 shows the distribution of votes per resolve is heavy tailed, although not nearly as skewed as the votes per user. The points are approximately a lognormal distribution, which arises from a decaying multiplicative reinforcement mechanism where resolves accrue votes at a rate proportional to the number of votes they already have (Wu & Huberman 2007). Such a mechanism may arise for several reasons, including global popularity-based search (Limpert, Stahel, & Abbt 2001; Mitzenmacher 2004; Huberman & Adamic 1999) and network diffusion[3] (Leskovec, Adamic, & Huberman 2007).

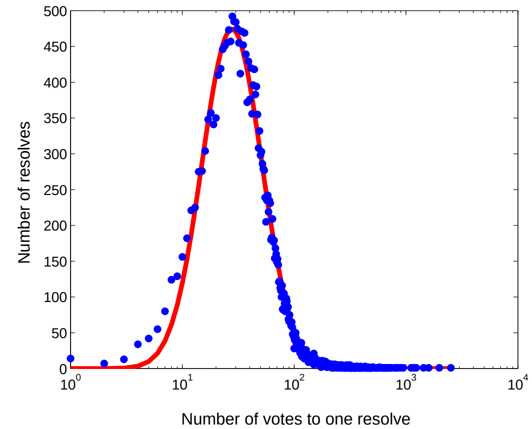

Figure 1: The frequency distribution of resolves vs. number of votes. The solid line indicates a lognormal fit to the data points, where the mean and standard deviation of the variate's logarithm are 3.7 and 0.62, respectively.

Networks and user participation

This section describes correlations between network participation and voting activity. 5,358 Essembly users (43%) participated in at least one network. The top of Table 1 lists the total nodes and links for each network. Users with network links were much more active than the others, accounting for 96% of all votes. The average number of votes per networked user was 286, versus 6 for non-networked users, a statistically significant difference[4].

Fig. 2 shows many users participated in more than one network. Most interested, dedicated users adopted Essembly's multiple relationship types: of the 622 users in the top 5% of voting activity, 85% were in at least two networks.

The voting activity section of Table 1 shows networked users have a positive Pearson correlation between voting and

[2] The data also includes the time and length of each comment posted for a resolve and the number of responses to that comment.

[3] Network diffusion assumes users of different degrees are well-mixed, as is the case in the Essembly networks as we later show.

[4] Wilcoxon rank sum test, $p < 10^{-12}$

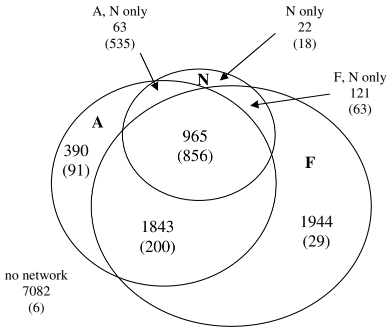

Figure 2: Number of users, and their average numbers of votes in parentheses, who are in each combination of networks. Labels F, A and N denote the friends, allies and nemeses networks, respectively.

degree in all three networks. Thus not only are networked users more active than non-networked, we find dedicated users tended to use both the networking and voting features.

Network structure

This section presents the structure of the three Essembly networks. Each network has a heavy-tailed degree distribution, which is fairly well approximated by a truncated power law: the fraction of nodes of degree d is proportional to $d^{-\tau} e^{-d/\kappa}$ except the top percent of degrees are somewhat larger than this fit suggests. Table 1 quantifies the degree distributions. Such a skewed distribution is typical of online social networks (Newman 2003). The truncated power law can arise from a combination of preferential attachment and users becoming inactive (Amaral *et al.* 2000).

The three networks contain few common links:

$$P(\text{allies}|\text{friends}) = 29\% \qquad P(\text{friends}|\text{allies}) = 25\%$$
$$P(\text{nemeses}|\text{friends}) = 3\% \quad P(\text{friends}|\text{nemeses}) = 20\%$$

This reflects one purpose of the ideological networks in Essembly, namely to help users find others with similar views beyond their friends[5]. This observation is consistent with users creating links with different nominal semantics.

Each network has a giant connected component containing all the high degree nodes and most low degree nodes. Thus most nodes, including all the most highly connected, are reachable from most other nodes by following network links. Large online social networks that have been studied have a giant component (Newman 2003).

The networks differ in structure within their giant components, as indicated in the community structure section of Table 1. One example is the difference in transitivity, a measure of local clustering in a graph given by the average over the nodes of the ratio of existing links to possible links among each user's first neighbors (Watts & Strogatz 1998).

[5]A few (122) alliances are also nemeses, but most of these (88) are also friends. So the majority of users seem to see allies and nemeses as mutually exclusive. We discarded the small minority who didn't (less than 1% of all relationships) in the homophily plots and decision tree analysis below.

network	friends	allies	nemeses
nodes	4873	3261	1181
links	13516	15593	2075
community structure			
transitivity	0.31	0.19	0.02
modularity	0.73	0.47	0.56
average path length	4.67	3.77	4.37
degree distribution			
power law tail, τ	1.2	1.2	1.4
cutoff, κ	27	57	18
median	2	3	2
maximum	274	382	125
degree assortativity	−0.02	−0.11	−0.15
components			
giant component fraction	0.87	0.96	0.85
number of components	229	59	77
voting activity			
correlation with degree	0.28	0.51	0.34
voting assortativity	0.22	0.06	0.01

Table 1: Properties of the three Essembly networks. The quantities are described in more detail in the text.

Social networks typically have values from 0.3 to 0.8 (Newman 2003). A broader measure of community structure is modularity which characterizes the extent to which the network clusters into well-defined communities, with reported values for social networks in the range $0.4 - 0.9$ (Newman 2006). The values in Table 1 are estimates with a greedy algorithm for community structure (Clauset 2005). Both modularity and transitivity are higher for the friends network than for the other two networks, with especially small transitivity among nemeses. The friends network is thus moderately clustered, while users in the allies and nemeses networks have less well-defined communities.

To check these conclusions on community structure differences, we compared the differences in these values between each pair of networks with what one would expect under the hypothesis that users selected the type of link for each network in the pair randomly but in the same proportion as in the actual networks. This randomization test (Cohen 1995) showed the differences we observed are unlikely (with p-value less than 5% in all cases) to arise with random link type choices.

Another important difference in the networks is the average path length, the average shortest distance between all pairs of nodes in the giant component. The path lengths reported in Table 1 show the networks have the small world property typical of social networks (Newman 2003). The relatively high average path length in the friends network is related to its high modularity. Users tend to form closely-knit communities in the friends network, so that one or two inter-community links are often required to connect pairs of nodes. By contrast, the allies network is more dense and star-like, with most users directly connected to the core of high degree nodes.

Taken together, these observations indicate a friends network typical of online social networks, a star-like allies network with structure similar to the Internet (on a much smaller scale), and a sparse nemeses network.

We gain further insight into the networks by comparing how links relate to two measures of activity in Essembly: forming links and voting. Degree and voting assortativity measure the tendency of people with similar levels of those activities to link to each other. Table 1 shows Essembly's degree assortativity is negative and lower than that of other online social networks (Newman 2003) whereas the voting assortativity is positive. Thus users with many neighbors tend to link to people with few neighbors, while active voters tend to link to other active voters. One explanation for this contrast is that Essembly facilitates finding other users through their votes rather than only their networks.

User homophily and network distance

Friends tend to display a certain level of homophily, that is, similarity of opinion or interest (McPherson, Smith-Lovin, & Cook 2001). Homophily between Essembly users can be estimated by comparing the votes they make on resolves in common. In this section we compare user homophily to distance in the networks, which suggests users are faithful to link semantics when making connections.

Essembly calculates homophily and allows members to compare themselves to other users whether they are network neighbors or not. In light of the four-point voting scale, Essembly uses the following L_1 metric[6], which we also adopt:

$$S(a,b) = 1 - \frac{1}{|R_{a,b}|} \sum_{r \in R_{a,b}} \frac{1}{3} \left(|V(a,r) - V(b,r)| \right) \quad (1)$$

where $V(a,r)$ is the numerical value of the vote of user a on resolve r and $R_{a,b}$ is the set of resolves both a and b have voted on[7]. There is a moderate overall tendency to agree: the median homophily over all networked users is 0.64.

Fig. 3 shows homophily levels for first neighbors in the three networks, as well as randomly selected pairs. Friends and allies tend to be more ideologically similar (medians 0.71 and 0.77, respectively), while nemeses are decidedly dissimilar (median 0.43). People are more likely to agree with their allies than their friends, and with their friends than their nemeses. This provides evidence that the three semantically distinct networks capture different levels of ideological homophily.

Fig. 4 shows how homophily varies with distance in the networks and suggests that the links correspond to their nominal interpretation. The behavior for the nemeses network is particularly striking, indicating that, in Essembly, the enemy of my enemy could be my friend or ally. The oscillations in the nemeses curve show this behavior continues far into the network, in spite of averaging over many users.

[6]Chris Chan, personal communication

[7]We include the ten initial resolves in this analysis. Essembly does not include user-designated "just for fun" resolves in its calculation, but this distinction was not made in our data so we included all resolves.

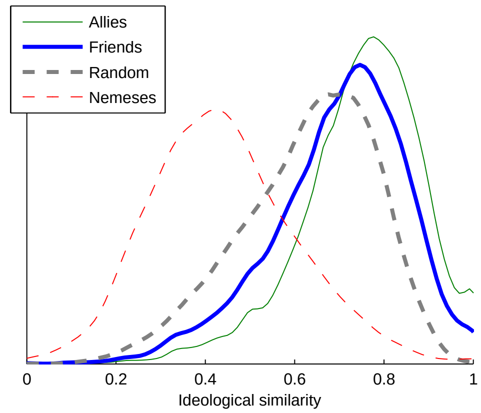

Figure 3: Distribution of vote similarities for pairs of users who have voted on at least 10 resolves in common. Curves are for neighbors in the three networks, as well as random pairs of users.

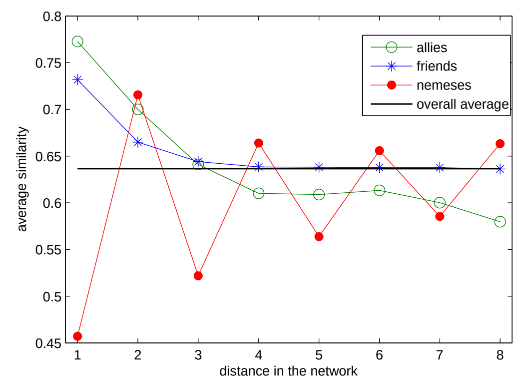

Figure 4: Vote similarities vs. distance in friends, allies and nemeses networks (blue, green and red curves, respectively). The curves use the similarity measure of Eq. (1).

Network effect on voting

A basic application of a social network in an online service is to spread information about site content. In this section we examine the role of the Essembly networks in encouraging voting. We find a moderate overall effect but strong effects for certain combinations of factors.

One way to measure the network effect is to count the number of votes cast by users before any of their network neighbors have voted. These votes are certainly not inspired by a neighbor's vote and we will refer to them as *seeds*. The percentage of seeds is not a useful measure by itself, because it is impossible to tell (given our data) whether a non-seed resulted from network influence.

To establish the significance of the percentage seed votes, we performed a randomization test where the actual voting histories were compared to simulated histories where users "voted" on resolves with no regard to the network structure.

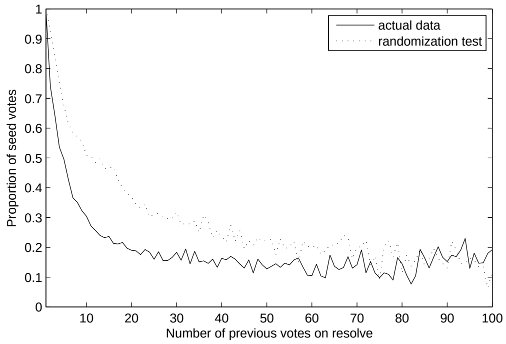

Figure 5: Percentage of seed votes (votes cast prior to any network neighbor voting) as a function of the number of previous votes. The percentages are averages over resolves. Standard deviations (not pictured) are on the order of 0.01 for all data points.

In these simulated histories, users were constrained to vote with the same frequency as in the real data, and resolves constrained to accumulate votes at the real rate, but beyond these constraints users chose resolves at random[8].

Fig. 5 shows the difference in percentage of seed votes for the actual history compared to the average of a number of randomization tests. We see this difference regardless of the number of votes a resolve received. The comparison of seed vote percentage shows a statistically significant, if moderate, network effect. Because the networks are rather dense, with highly active users also having many network connections, the subnetwork of voters on a given resolve quickly percolated throughout the overall network. That is, a giant connected component of voters formed in the real data and the randomization tests, after only a few votes. Indeed, the real data and the randomization tests had statistically indistinguishable percentage of voters in the giant component for resolves with more than 50 votes. This result highlights the importance of high-degree nodes in the process of spreading new content, no matter what the mechanism.

On the other hand, in a few cases where relatively large numbers of network neighbors voted previously, we find significant correlations using a decision tree analysis (Brzozowski, Hogg, & Szabo 2008). We considered all 1.3 million votes. To identify when networks may have been particularly influential we needed examples where users could have voted on a resolve but didn't. We created these examples by inferring user sessions: periods of activity for a given user, punctuated by three or more hours without any votes.

[8]These tests used only the 2000 resolves created during the month ending Nov. 22, 2006, giving both nearly complete networks and sufficient time to accumulate votes. Specifically, these resolves are from near the end of our data sample, so the networks were likely almost all established before these resolves. Furthermore, these resolves were created at least 20 days before the end of our sample, on Dec. 12, 2006, giving enough time to accumulate most of their votes: on average, a resolve received 93.6% of its votes in the first 20 days after its introduction.

friends	allies	nemeses	Posterior P_{vote}	N
0	0	0	44%	804,299
1+	0	0	65%	96,910
4+	0	0	73%	3,022
0	0	2+	38%	12,462
0	1+	0	49%	604,042
0	1+		49%	825,656

Table 2: Posterior probability of voting, as a function of how many friends, allies, and nemeses voted previously. N is the number of samples for each of the cases. The last row corresponds to having at least one ally or nemeses. (The prior probability is set at P_{vote} = 50%.)

At the end of each session, for each vote a user placed during the session, we created one negative example by randomly selecting a resolve the user could have voted on (from the set of all resolves that existed at the time) but didn't. This gives roughly equal numbers of positive and negative examples.

We learned a decision tree over the complete set of examples using (Chickering 2002). The leaves reflect the portion of cases where a user voted under the given conditions, relative to the prior distribution ($P_{\text{vote}} = 1/2$). As features we used the number of friends, allies, and nemeses who'd voted previously and thus could have shown up on a user's home page as an influence.

Table 2 shows some results from the leaves of the tree. Rows 1 and 2 suggest users are 48% (i.e., $65/44 - 1 = 0.48$) more likely to vote on a resolve if one or more of their friends voted on it than if none of their neighbors did ($p < 0.001$). By contrast, if four or more friends (and no other neighbors) voted on a resolve already, users are 66% more likely to vote on it (Table 2 rows 1-3). Curiously, when two or more nemeses voted on a resolve (but no friends or allies), the posterior P_{vote} drops, to 38%: a user is 14% less likely to vote on a resolve if two or more nemeses voted on it than if no one in his or her local networks did ($p < 0.001$; Table 2 rows 1, 4).

Essembly encourages users to vote on resolves voted on by their allies, and there are indeed many examples where a user sees only allies voting on a resolve. However, users are not significantly more likely to vote on resolves their allies have voted on (Table 2 row 5). Users are also 33% more likely to vote on a resolve that friends voted on than some combination of allies and nemeses ($p < 0.001$; Table 2 rows 2, 6). This suggests that friends are more influential than allies or nemeses in selecting resolves.

Discussion

We examined user activity and network connections from Essembly, an online service for people interested in political discussion and organization. Essembly provides three different relationship types, creating three separate networks of the same users. Our results show Essembly is a useful test case of an online social network allowing multiple relationship types. Furthermore, Essembly users follow typical patterns of activity and social network structure, so these observations may generalize to other online communities.

Users were faithful to the link semantics and the networks had different structural and functional properties, suggesting that distinct networks may have useful applications. For example, friends connections are most relevant to applications related to communities and persuasion. The allies network is well-suited to the rapid spread of information and forming significant collaborations. Our observation that nemesis activity inhibits voting suggests that a "nemeses" network could function as an ignore list, possibly overlaid on a friends network, for people to identify users whose content choices are not likely to be interesting. The high homophily demonstrated by enemies of enemies suggests the nemesis network could also identify potential allies. This behavior is analogous to the propagation of trust and distrust in networks (Guha *et al.* 2004).

The networks played a surprisingly moderate role in spreading new content, with network connections becoming more important as resolves age. This contrasts with other web sites, such as Digg (Lerman 2007), where new content is not shown on the main page until it achieves a certain level of popularity and early votes are mostly within the networks. This suggests that novelty and overall popularity are powerful means of spreading content even in the presence of a social network. These mechanisms could be included in our analysis as resolve features, such as its age or overall popularity, independent of network neighbors.

It would be interesting to model voting behavior based on the approaches the Essembly user interface provides for finding resolves, *e.g.*, recent, popular, controversial or active among user's network neighbors. It would also be useful to compare multiple network types on the same set of users in other situations, especially with records of link creation times, and perform intervention experiments to help identify which correlations in behavior are causal (*e.g.*, by changing types of links available to different subgroups of users).

Online social networks allow incorporating relationship information into web services and economic mechanism designs. When it is useful to match semantics of a link to its use, Essembly shows it is feasible to have users explicitly specify multiple link types relevant to their task. Distinguishing homophilous or other attributes of interest from friendship adds meaning to each, in particular, easing the burden on the socially loaded concept of friendship.

Acknowledgments

We thank Chris Chan and Jimmy Kittiyachavalit of Essembly for their help in accessing the Essembly data. We thank Scott Golder for helpful comments.

References

Adamic, L. A.; Buyukkokten, O.; and Adar, E. 2003. A social network caught in the web. *First Monday* 8(6).

Ahn, Y.-Y., et al. 2007. Analysis of topological characteristics of huge online social networking services. In *Proc. of 2007 Intl. Conf. on the World Wide Web*, 835–844.

Amaral, L. A. N.; Scala, A.; Barthelemy, M.; and Stanley, H. E. 2000. Classes of small-world networks. *Proc. of the Natl. Acad. Sci.* 97:11149–11152.

boyd, d. 2006. Friends, friendsters, and top 8: Writing community into being on social network sites. *First Monday* 11(12).

Brzozowski, M.; Hogg, T.; and Szabo, G. 2008. Friends and foes: Ideological social networking. In *Proc. of the SIGCHI Conference on Human Factors in Computing (CHI2008)*. ACM Press.

Chickering, D. M. 2002. The WinMine toolkit. Technical Report MSR-TR-2002-103, Microsoft.

Clauset, A. 2005. Finding local community structure in networks. *Physical Review E* 72:026132.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

Gladwell, M. 2000. *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Little Brown.

Goldberg, D.; Nichols, D.; Oki, B. M.; and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM* 35(12):61–70.

Golder, S. A.; Wilkinson, D. M.; and Huberman, B. A. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Proc. of the 3rd Intl. Conf. on Communities and Technologies (CT2007)*.

Guha, R.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2004. Propagation of trust and distrust. In *Proc. of the 13th Intl. World Wide Web Conf. (WWW2004)*, 403–412. New York: ACM.

Hogg, T., and Adamic, L. 2004. Enhancing reputation mechanisms via online social networks. In *Proc. of the 5th ACM Conference on Electronic Commerce (EC'04)*, 236–237. ACM Press.

Huberman, B. A., and Adamic, L. A. 1999. Growth dynamics of the World Wide Web. *Nature* 401:131.

Lam, C. 2004. SNACK: incorporating social network information in automated collaborative filtering. In *Proc. of the 5th ACM Conference on Electronic Commerce (EC'04)*, 254–255. ACM Press.

Lerman, K. 2007. Social information processing in social news aggregation. preprint arxiv.org/abs/cs.cy/0703087.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Transactions on the Web* 1(1).

Limpert, E.; Stahel, W. A.; and Abbt, M. 2001. Log-normal distributions across the sciences: Keys and clues. *BioScience* 51:341–352.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Mitzenmacher, M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1:226–251.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review* 45(2):167–256.

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* 103:8577–8582.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM* 40(3):56–58.

Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440–442.

Watts, D. J.; Dodds, P. S.; and Newman, M. E. J. 2002. Identity and search in social networks. *Science* 296:1302–1305.

Wilkinson, D. M., and Huberman, B. A. 2007. Assessing the value of cooperation in Wikipedia. *First Monday* 12(4).

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proc. of the Natl. Acad. Sci.* 104:17599–17601.