

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

3-2010

Homophily in the Digital World: A LiveJournal Case Study

Hady W. LAUW

Singapore Management University, hadywlawu@smu.edu.sg

John C. SHAFER

Microsoft Research

Rakesh AGRAWAL


Microsoft Research

Alexandros NTOULAS

Microsoft Research

DOI: <https://doi.org/10.1109/MIC.2010.25>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Citation

LAUW, Hady W.; SHAFER, John C.; AGRAWAL, Rakesh; and NTOULAS, Alexandros. Homophily in the Digital World: A LiveJournal Case Study. (2010). *IEEE Internet Computing*. 14, (2), 15-23. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/1514

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.



Homophily in the Digital World

A LiveJournal Case Study

Are two users more likely to be friends if they share common interests? Are two users more likely to share common interests if they're friends? The authors study the phenomenon of homophily in the digital world by answering these central questions. Unlike the physical world, the digital world doesn't impose any geographic or organizational constraints on friendships. So, although online friends might share common interests, a priori there's no reason to believe that two users with common interests are more likely to be friends. Using data from LiveJournal, the authors show that the answer to both questions is yes.

A common assumption about human nature is that people have a tendency to associate with other, similar people (a phenomenon called *homophily*). Sociology has studied homophily in the physical world extensively. As surveyed elsewhere,¹ researchers have examined its effects on many types of real-world relationships, ranging from marriage to casual acquaintance. However, the studies have generally been conducted on a small scale, and the similarity factors examined have been limited mostly to easily observed or surveyed sociodemographic characteristics, such as race, gender, religion, and occupation – characteristics that don't necessarily manifest themselves in online social networks.

One of the strongest underlying sources of homophily in the physical world is locality due to geographic

proximity, family ties, and organizational factors, such as school and work.¹ However, in the digital world, physical locality becomes less important, and other factors such as common interests might play a greater role. Here, we look at two central questions regarding homophily in the digital world:

- Are two users more likely to be friends if they share common interests?
- Are two users more likely to share common interests if they're friends?

Although these questions are related, we posit that the former is the more interesting one. Although it might be reasonable that friends would have common interests, no a priori basis exists to think that within a large and diverse network, two arbitrary users

Hady W. Lauw,
John C. Shafer,
Rakesh Agrawal,
and Alexandros Ntoulas
Search Labs, Microsoft Research



Figure 1. LiveJournal User Info page for *reis_gym*. The page lists the user’s friends, interests, and communities.

with a common interest are any more likely to be friends, especially when they aren’t confined by geographic or organizational constraints. What’s the likelihood that two users from different continents would become friends if they have the same hobbies?

To answer these questions, we analyzed numerous User Info pages collected from LiveJournal (www.livejournal.com), a popular blogging and social networking platform. LiveJournal users identify each other as friends and express their interests in two ways. First, each user has a list of self-proclaimed interests on his or her User Info page. Second, users can subscribe to communities or group blogs oriented around a given topic. This presence of friendship links and expression of interests makes LiveJournal an appropriate experimental vehicle to investigate the questions we’re asking. Moreover, because we observed that some users have several friends and interests, whereas others have few, we conducted our study across groups with varying involvement levels.

Data Sets

The data we use in this study consists of LiveJournal users’ lists of friends, interests, and communities as specified on their User Info pages. We chose data on the basis of its public availability. Figure 1 illustrates the User Info page for user *reis_gym* (www.users.livejournal.com/reis_gym/profile). Friendship in LiveJournal has at least two connotations: those whom the user cares to hear from and follow, and those whom the user would trust with more sensitive information. These connotations arise from the role of friendship as a distribution and privacy tool. LiveJournal automati-

cally notifies users of their friends’ most recent blog entries. Meanwhile, a blog entry could be marked “friends only” to limit its distribution to only friends. Although reciprocation isn’t necessary for “friending” someone, it bolsters the confidence in the friendship claim. So, in this article, we consider two users friends only if they’ve mutually friended each other.

Users can freely enter any word or short phrase as an interest. For example, *reis_gym* listed interests such as cycling, martial arts, running, and weightlifting (see Figure 1). LiveJournal also hosts group blogs or communities. “Watching” a community has an effect similar to friending a user, in that users are notified of blog entries the community posts. So, watching a community is another form of expressing an interest. For instance, *reis_gym* was watching *ddr_exercise*, a community about Dance Dance Revolution (DDR) exercise, and *exercisupport*, a community for people trying to get in shape.

Graphs Capturing Friendship, Interests, and Communities

Starting from a random seed set, we collected User Info pages (roughly 309,000) by crawling LiveJournal over several days in January 2009. Although LiveJournal has more than 20 million user accounts, not all of them are active (www.livejournal.com/stats). To estimate the fraction of active users our data covers, we used a LiveJournal feature that returns the blog page of a random user (www.livejournal.com/random.bml) who was active in some way. We thus obtained a random sample of 5,000 active users. By computing the number of pages from this set that were absent in the crawl set, we estimated that our data set covered slightly more than one-fifth of active users.

We then extracted the following three graphs, expressed as binary adjacency matrices:

- **F** is the user \times user friendship graph. $F_{uu'} = 1$ if users u and u' have friended each other, and 0 otherwise. By definition, friendship is symmetric – that is, $\forall u' \neq u, F_{uu'} = F_{u'u}$ – but not reflexive – that is, $\forall u, F_{uu} = 0$.
- **I** is the user \times interest graph. $I_{ui} = 1$ if user u specifies i as an interest, and 0 otherwise.
- **C** is the user \times community graph, where $C_{uc} = 1$ if user u watches community c , and 0 otherwise.

Table 1. Dimensionality and density of the friendship, interest, and community graphs.

	Friendship F	Interest I	Community C
Dimension	263,838 users × 263,838 users	172,472 users × 539,707 interests	17,496 users × 78,761 communities
Density (%)	0.009	0.006	0.02

Table 2. Data sets based on involvement levels.

	Users	Interests	Communities
All	12,451	140,485	70,388
Active	1,024	31,210	52,505
Highly Active	161	11,919	17,081

Table 1 shows these graphs' sizes after removing nodes without any links. Of the more than 300,000 User Info pages in our crawl, 263,000 (85 percent) have at least one friend, 172,000 (56 percent) have at least one interest, and 17,000 (6 percent) have at least one community. It's reasonable that fewer users have communities than interests because watching communities involves a deeper commitment than simply stating interests.

Table 1 also shows these graphs' densities, defined as the fraction of existent links over the total number of possible links. All three are rather sparse graphs, with densities ranging from 0.006 percent to 0.02 percent. This sparsity is expected because any one user can keep up with only a certain number of friends, interests, or communities. These graphs are also well connected: the single largest connected component covers more than 99 percent of the users in F and I and 90 percent of those in C. This level of connectivity suggests a small-world network structure.²

User Involvement

We observed that more information existed about some users than others. This, however, doesn't necessarily imply that users with less information are less social; they might simply be less involved with LiveJournal. So, it's useful to divide the users into groups of various involvement levels to see whether similar results hold for these different groups. We measure involvement as the minimum number of links that a user has in each of the three graphs F, I, and C. For subsequent analysis, we use three data sets of different involvement levels as follows:

- All contains the users who have at least

one friend, interest, and community; that is, $\sum_{u'zu} F_{uu'} \geq 1$, $\sum_i I_{ui} \geq 1$, and $\sum_c C_{uc} \geq 1$.

- Active contains the users who have at least 10 friends, interests, and communities; that is, $\sum_{u'zu} F_{uu'} \geq 10$, $\sum_i I_{ui} \geq 10$, and $\sum_c C_{uc} \geq 10$.
- Highly Active contains the users who have at least 50 friends, interests, and communities; that is, $\sum_{u'zu} F_{uu'} \geq 50$, $\sum_i I_{ui} \geq 50$, and $\sum_c C_{uc} \geq 50$.

Table 2 shows the data sets' sizes. All contains all the users who exist in all three graphs, with 12,451 users having 140,485 interests and 70,388 communities. As we increase the level of involvement, the number of users reduces dramatically. The most active subset, Highly Active, has 161 users, 11,919 interests, and 17,081 communities.

Friendship and Interests

Here, we address the two central questions we raise in this article: whether having common interests makes it more likely for a pair of users to be friends, and whether being friends influences the likelihood of having common interests. Our methodology for answering the first question is to see whether a significant increase occurs from the prior probability of friendship to the conditional probability of friendship given that a pair of users has common interests. Similarly, for the second question, we compare the prior probability of having common interests with the conditional probability given that the pair is composed of friends.

Probability of Friendship

In a large, diverse network, we normally wouldn't expect that two arbitrary people with common

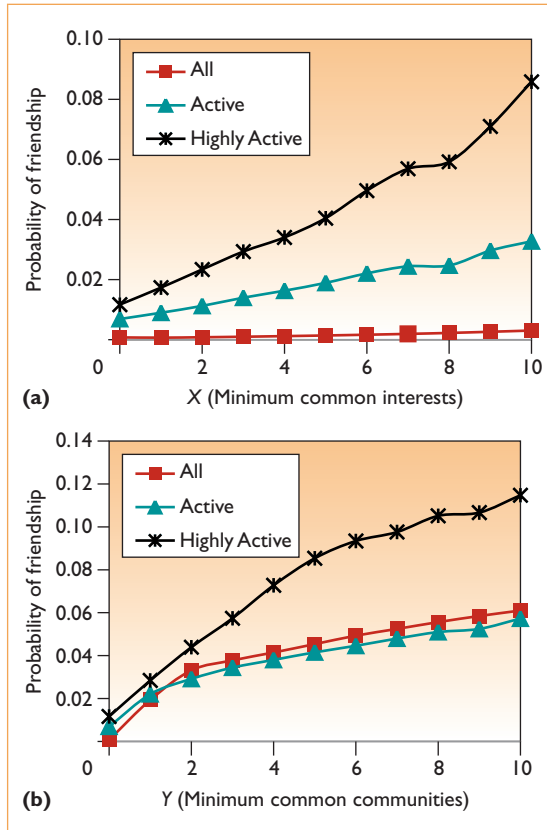


Figure 2. The probability of friendship. We calculated the probability that a pair of users would be friends given (a) common interests and (b) common communities.

interests are more likely to be friends. However, the Web transcends geography and other organizational foci, and interest might play a more significant role as a conduit for people to get to know and interact with each other. We can see whether this holds up in the LiveJournal data.

Without any prior information, the best estimate for the probability of friendship $P(\text{friendship})$ is the fraction of random pairs that turn out to be friends. This is equivalent to the density of the friendship graph F computed by Equation 1, where U denotes the set of users in consideration:

$$P(\text{friendship}) = \frac{\sum_{u \in U} \sum_{(u' \neq u) \in U} F_{uu'}}{|U| \times (|U| - 1)}. \quad (1)$$

Conditional on common interests. Given that a user pair shares a minimum number of X common interests, we can compute the conditional probability of friendship as follows. Let I_u be the vector representation of the u th row of I .

The dot product $I_u \cdot I_{u'}$ gives the number of common interests for the user pair (u, u') . The probability of friendship given X or $P(\text{friendship} | X)$ is then the fraction of user pairs with at least X common interests who turn out to be friends, as Equation 2 computes:

$$P(\text{friendship} | X) = \frac{|\{(u, u') \in U \times U | (F_{uu'} = 1) \wedge (I_u \cdot I_{u'} \geq X)\}|}{|\{(u, u') \in U \times U | (u \neq u') \wedge (I_u \cdot I_{u'} \geq X)\}|}. \quad (2)$$

Figure 2a plots $P(\text{friendship} | X)$ for different values of X and different data sets. $P(\text{friendship} | X = 0)$ is equivalent to the probability of friendship at random, $P(\text{friendship})$. Figure 2a shows that having common interests, even just one ($X = 1$), significantly increases the probability of friendship for all data sets. This trend is also monotonic: higher X leads to higher probability. This observation stands across data sets for various involvement levels.

This is a surprising outcome, given that without geographic constraint, we wouldn't expect the conditional probability to be significantly higher. It suggests that an underlying factor is at work in LiveJournal that encourages users to make friends with those having common interests. Several LiveJournal features might contribute to this. For every interest with more than one claimant, LiveJournal provides a hyperlink to the list of users who claim that interest, thus letting one user find others to connect with on the basis of interest. Blogging and commenting are another set of activities that could help users get to know others who share similar interests.

Conditional on common communities. We now investigate whether a similar relationship exists between friendship and common communities. Equation 3 computes the probability of friendship given that a user pair shares a minimum of Y common communities:

$$P(\text{friendship} | Y) = \frac{|\{(u, u') \in U \times U | (F_{uu'} = 1) \wedge (C_u \cdot C_{u'} \geq Y)\}|}{|\{(u, u') \in U \times U | (u \neq u') \wedge (C_u \cdot C_{u'} \geq Y)\}|}. \quad (3)$$

Figure 2b plots $P(\text{friendship} | Y)$ for different Y values and data sets. We observe similar trends as those in Figure 2a: a user pair is

monotonically more likely to consist of friends if they share more common communities.

To test these results' statistical significance, we compare the conditional probability computed on the original graphs with the respective probability computed on randomized graphs that maintain the same structural properties as the originals. We used swap randomization³ to create 100 random graphs while maintaining the same number of edges as in the I and C graphs. For the hypothesis that the conditional probability of friendship given common interests or communities is higher than the prior probability, our result is statistically significant with greater than 99 percent confidence. This statistical significance also holds for the stronger hypothesis involving a different randomization that preserves the node degrees in the original graphs, in addition to preserving the total number of edges. It's thus extremely unlikely that the higher conditional probability of friendship is due to the graph's structural properties.

Probability of Common Interests or Communities

Given that two users are friends, it's reasonable that friendship would increase the likelihood of common interests because when friends pursue activities together, they're likely to converge on such common interests.

Common interests. Equation 4 gives the probability that a random user pair shares at least X common interests $P(X)$:

$$P(X) = \frac{|\{(u, u') \in U \times U | (u \neq u') \wedge (I_u \cdot I_{u'} \geq X)\}|}{|U| \times (|U| - 1)}. \quad (4)$$

We compute the probability $P(X | \text{friendship})$ over only the set of pairs who are friends, as Equation 5 shows:

$$P(X | \text{friendship}) = \frac{|\{(u, u') \in U \times U | (F_{uu'} = 1) \wedge (I_u \cdot I_{u'} \geq X)\}|}{\sum_{u \in U} \sum_{u' \in U} F_{uu'}}. \quad (5)$$

Figure 3a compares $P(X)$ (darkly shaded) versus $P(X | \text{friendship})$ (lightly shaded) for different X values on the Highly Active data set. It shows that for every X , $P(X | \text{friendship})$

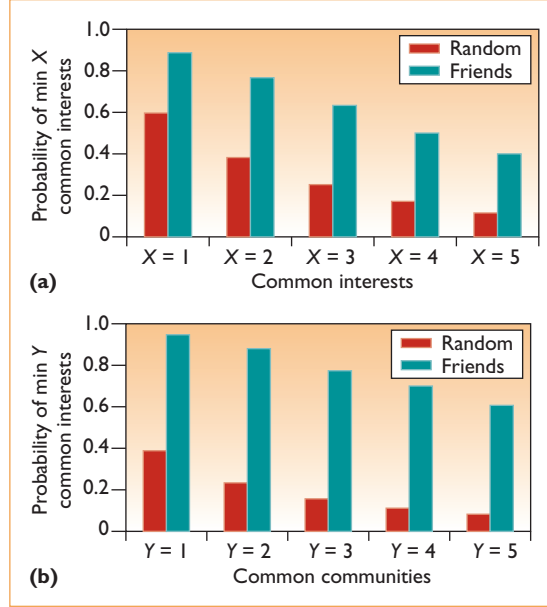


Figure 3. The probability of commonality given friendship. Using the *Highly Active* data set, we calculated the probability for (a) common interests and (b) common communities.

is significantly higher – between 1.5 and 3.5 times higher – than $P(X)$. The likelihood of common interests conditioned on friendship is as high as $P(X = 1 | \text{friendship}) = 0.89$ and $P(X = 2 | \text{friendship}) = 0.77$. This result suggests that friendship is a potentially significant source of signals in inferring a person's interests.

Common communities. We can conduct a similar exercise on communities. Figure 3b plots $P(Y)$ (see Equation 6) and $P(Y | \text{friendship})$ (see Equation 7) for various Y 's and for the Highly Active data set:

$$P(Y) = \frac{|\{(u, u') \in U \times U | (u \neq u') \wedge (C_u \cdot C_{u'} \geq Y)\}|}{|U| \times (|U| - 1)}. \quad (6)$$

$$P(Y | \text{friendship}) = \frac{|\{(u, u') \in U \times U | (F_{uu'} = 1) \wedge (C_u \cdot C_{u'} \geq Y)\}|}{\sum_{u \in U} \sum_{u' \in U} F_{uu'}}. \quad (7)$$

We see similar trends as in Figure 3a, but the difference is even higher. $P(Y | \text{friendship})$ is 2.4 to 7.3 times higher than $P(Y)$, suggesting that friendship is an even stronger signal in detect-

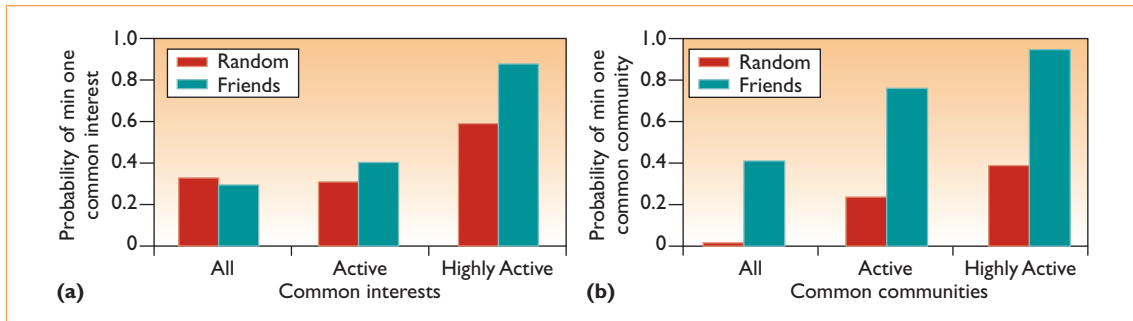


Figure 4. The probability of commonality for different activity levels. We calculated this probability for pairs of users with (a) at least one common interest and (b) at least one common community.

ing common communities. Figure 4a shows $P(X = 1 \mid \text{friendship})$, and Figure 4b shows $P(Y = 1 \mid \text{friendship})$ for different data sets of various involvement levels. Notably, the trend whereby friendship improves the probability of common interests and communities generally holds across these data sets.

A similar test we described earlier reveals that the higher conditional probability of common interests or communities given friendship is statistically significant.

Strength of Friendship

In addition to binary friendship, we're interested in studying friendship strength's effect on common interests or communities. Although some social network platforms, such as Orkut, let users indicate friendship strength, LiveJournal doesn't have this feature. So, we determine friendship strength based on the probability that a random walk starting from one user in a friendship pair would reach the other user on the friendship graph. Our methodology to study friendship strength's effect on interest is to measure the Pearson's correlation between friendship strength and the fraction of interests that a user pair has in common. A larger positive correlation value will indicate that a user tends to share more interests with stronger friends as compared to weaker ones.

Estimating Friendship Strength

We estimate the friendship strength between a user u and other users by performing a random walk with restart⁴ on the friendship graph, starting (and periodically restarting) from u . For a given user u , we compute the friendship-strength vector R_u by solving Equation 8:

$$R_u = (1 - \lambda) \cdot A \cdot R_u + \lambda \cdot \mathbf{1}(u). \quad (8)$$

Here, $\mathbf{1}(u)$ is a vector having element u set to 1 and all others to 0, A is the friendship graph's adjacency matrix, and λ is the damping factor denoting the likelihood of a restart from u during our random walk. We normalize entries for each user u in A using the number of outgoing links from u to his or her friends. We use a default value of $\lambda = 0.85$.⁴

The solution R_u from Equation 8 captures the probability of reaching other users by starting from user u in the friendship graph; we thus use it as a proxy to estimate friendship strength among users. Researchers used a similar random walk method elsewhere⁵ to estimate directional trust relationships. We can now define a weighted version of the friendship matrix, denoted F_w , to be a $|U| \times |U|$ matrix, where element (u, u') is set to the average of $R_u(u')$ and $R_{u'}(u)$.

To distinguish it from the weighted F_w , we use F_b to denote the binary matrix F we introduced previously. Similarly to F_b , F_w is symmetric but not reflexive. Given these two friendship-strength matrices, namely F_b and F_w , we can now study the correlation between friendship, interests, and communities.

Correlation with Interests and Communities

To measure the extent to which two users u and u' share common interests, we compute the Jaccard similarity between their sets of interests $I(u)$ and $I(u')$ – that is, $|I(u) \cap I(u')| / |I(u) \cup I(u')|$. Π^T denotes a vector in which each element corresponds to a pair of users and contains as a value the fraction of common interests between this pair. We compute the vector for communities CC^T similarly.

We then use the Pearson's correlation to measure the correlation between friendship strength and common interests or communities. For two series of values P and Q , we compute the Pearson's correlation using Equation 9:

Related Studies on User Behavior in Social Networks

Many researchers are interested in how social relationships might affect various user behaviors. David Crandall and his colleagues looked at social influence,¹ which is the degree to which relationships induce similarity, and selection, or the degree to which similarity induces relationships. The authors examined both Wikipedia and LiveJournal data and used activities (edits to Wikipedia articles and community membership in LiveJournal) as the basis for similarity. However, they didn't directly study friendship and interests as we do in the main text. Instead, they compared the ability of friendship and similarity to predict future activities — for example, the probability of joining a community in LiveJournal. Their conclusion showed that LiveJournal users were more likely to join communities if the most similar other users had already joined than if their friends had. However, this didn't mean that friendship had no predictive value. The probability of joining a community increased with the number of friends who had joined, which is in line with our conclusion that friendship affects interest similarity.

Other studies have focused on the relationship between friendship and interests in only one direction. For example, researchers studied the influence of friendship on similarity and showed that users who were friends on an instant-messaging platform were more likely to be similar in terms of topics of queries issued to a Web search engine as well as in demographic attributes, such as age, gender, and zip code.² Other research studied similarity's influence on friendship by analyzing user homepages.³ The authors modeled friendship in terms of hyperlinks between users' homepages and similarity in terms of common hyperlinks and homepages' textual content. These studies were conducted on approximations of interests (Web

queries) or friendship (hyperlinks). However, their conclusions still concur with ours, which are based on clearly specified interests and friendship.

With regard to estimating friendship strength in online communities, Eric Gilbert and Karrie Karahalios proposed a regression model for measuring friendship among Facebook users.⁴ They collected a relatively small amount of training data by surveying 35 users. The features they used for the model are related mainly to user interactions and demographics, which don't extend easily to data sets such as LiveJournal. A small category of features (called *structural features*) capture such variables as common groups and common words used in expressing interests. Without enumerating specific features, Gilbert and Karahalios showed that this category as a whole had a positive coefficient in the regression model, indirectly indicating that similarity in interests was useful in measuring tie strength.

References

1. D. Crandall et al., "Feedback Effects between Similarity and Social Influence in Online Communities," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD 08)*, ACM Press, 2008, pp. 160–168.
2. P. Singla and M. Richardson, "Yes, There Is a Correlation — from Social Networks to Personal Behavior on the Web," *Proc. World Wide Web Conf. (WWW 08)*, ACM Press, 2008, pp. 655–664.
3. L.A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Social Networks*, vol. 25, no. 3, 2003, pp. 211–230.
4. E. Gilbert and K. Karahalios, "Predicting Tie Strength with Social Media," *Proc. Int'l Conf. Human Factors in Computing Systems (CHI 09)*, ACM Press, 2009, pp. 211–220.

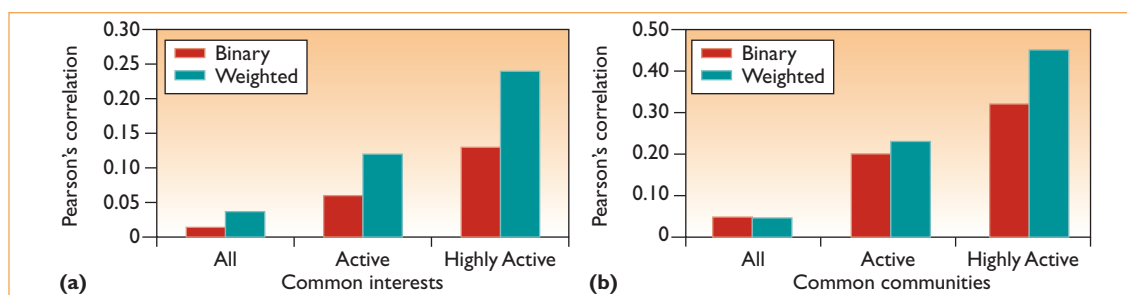


Figure 5. The correlation between common interests or communities and friendship. We can see the correlation between friendship strength and (a) common interests and (b) common communities.

$$r_{pq} = \frac{\sum (p_i - \bar{p})(q_i - \bar{q})}{(n-1)\sigma_p\sigma_q}, \quad (9)$$

where n is the cardinality of P and Q , and \bar{p}, \bar{q} are the means and σ_p, σ_q the standard deviations of P and Q , respectively. The correlation value ranges from -1 to 1 , with values closer to

1 or -1 indicating strong positive or negative correlation, whereas the value 0 shows no correlation at all.

Figure 5a shows the correlation between friendship strength and the common interests for different data sets. The x -axis represents the different data sets based on involve-

Table 3. Biclique examples.

Biclique 1	Biclique 2	Biclique 3	Biclique 4
html	avenue q	george harrison	dbsk
javascript	broadway	john lennon	j-pop
php	musicals	paul mccartney	k-pop
programming	new york city	peace	super junior
python	rent	ringo starr	tvxq
xml	wicked	—	—

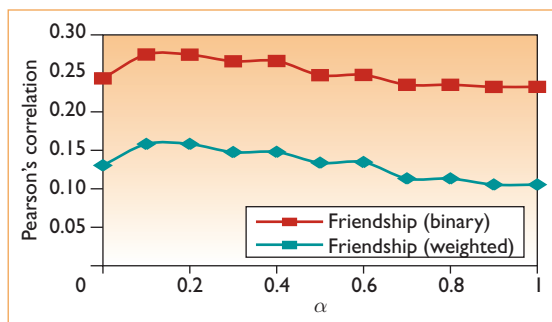


Figure 6. Correlation of friendship strength with common interests (*Highly Active*). We calculated the correlation between friendship strength and common interests for different values of α .

ment levels; the y-axis shows the correlation between friendship strength and the fraction of common interests. The darker bars correspond to the friendship strength captured by the binary version F_b , whereas the lighter bars depict the friendship strength from the weighted version F_w . We can see that, overall, the weighted friendship strength correlates better with the set of common interests. This implies that users' interests tend to overlap more when the weighted friendship indicates that they're stronger friends. In almost all cases, the correlation is twice that of the binary friendship.

Figure 5b shows the corresponding correlations between friendship strength and common communities. The results are similar to our previous ones: the weighted friendship strength correlates better than the binary one. On average, the correlation values with regard to communities are higher than those of interests. This could be because some interests might be noisy because users can specify anything as an interest, whereas this isn't true with communities, which a dedicated group of users more carefully creates and maintains.

Strength of Interests

Noisy interests – for example, *oh my god!*, *i miss you!*, and *lunch with messy people* – are generally random and unrelated to other valid interests, so we seek to remove them by finding groups of strongly related interests. Discovering these groups will also help us measure the strength of a user's association with a set of interests, given that users who have stronger association generally cite multiple related interests. We then use a correlation methodology similar to the one we described in the previous section to study how strength of interests affects friendship.

Finding Groups of Related Interests

As opposed to noisy interests, valid interests are usually related to other valid interests. As Figure 1 illustrates, user *reis_gym* used *aikido*, *jujutsu*, *kungfu*, and *muay thai* to express her interest in martial arts. This inspires us to find such valid interests by finding dense subgraphs in the bipartite user-interest graph I , in which several users have cocited numerous common interests. We define a dense subgraph as an (m, n) biclique⁶ – a maximally connected bipartite subgraph with at least m users and n interests.

Table 3 shows several examples of valid groups of related interests we discovered by finding $(8, 5)$ bicliques in the user-interest graph. Biclique 1 represents a general interest in Web programming and design, Biclique 2 in Broadway musicals, Biclique 3 in the Beatles, and Biclique 4 in Japanese and Korean pop music (DBSK, TVXQ, and Super Junior are well-known Korean pop bands).

Correlation with Friendship

We next studied how the strength of interests affects the correlation with friendship. To this end, we modeled the strength with which a user is associated with the set of interests in a biclique with $\alpha \in [0, 1]$. For a given α , we kept only the set of interests for which the user has at least α fraction of any biclique. For example, if $\alpha = 0.5$, a user must have at least three out of the six interests in Biclique 1 for us to associate them with that set of interests. This removes interests weakly associated with the user.

For different α values, Figure 6 plots the Pearson's correlation between the strength of

friendship (F_b and F_w) and the fraction of common interests (I^c), as we computed in the previous section. $\alpha = 0$ means that I includes all a user's interests, whereas $\alpha = 1$ means I includes only interests for which a user has the whole biclique. Although we conducted biclique discovery over the global graph, we did the correlation experiment for users in the `Highly Active` data set. We can make similar observations for other data sets.

The correlation initially increases from $\alpha = 0$ to $\alpha = 0.2$ and then monotonically decreases. The initial increase demonstrates the value of removing noisy interests not belonging to any biclique. However, it also suggests that a user doesn't need the whole biclique to have a strong interest in the topic it represents. Even a light association ($\alpha = 0.2$) is a sufficient indicator. Requiring a very high α is counterproductive because it could remove many interests that a user actually cares about. Importantly, the correlation with F_w is significantly higher than with F_b for all α .

This study of homophily using LiveJournal data shows that friendship and interests are strongly interlinked; having even a few common interests makes friendship significantly more likely. Equally important, being friends also makes a pair of users more likely to share common interests. It would be interesting to conduct a similar study on other social networks, such as Orkut or Facebook, assuming available data. Doing so would not only let us see how generally our conclusions hold in the digital world but would also enable a study of structural differences between the networks and how those might affect the role of interests in friendship. □

Acknowledgments

We thank Dennis Fetterly and Ming Ma for their help in the data collection, and Sreenivas Gollapudi, Nina Mishra, Panayiotis Tsaparas, and Raja Velu for helpful discussions.

References

1. M. McPherson, L. Smith-Lovin, and J.M. Cook, "Birds of a Feather: Homophily in Social Networks," *Ann. Rev. Sociology*, Aug. 2001, pp. 415–444.
2. D. Watts and S. Strogatz, "Collective Dynamics of Small-World Networks," *Nature*, vol. 393, no. 6684, 1998, pp. 440–442.
3. A. Gionis et al., "Assessing Data Mining Results via Swap Randomization," *ACM Trans. Knowledge Discovery from Data*, vol. 1, no. 3, 2007, art. 14.
4. G. Jeh and J. Widom, "Scaling Personalized Web Search," *Proc. World Wide Web Conf. (WWW 03)*, ACM Press, 2003, pp. 271–279.
5. M. Richardson, R. Agrawal, and P. Domingos, "Trust Management for the Semantic Web," *Proc. Int'l Semantic Web Conf.*, Springer, 2003, pp. 351–368.
6. S.P. Borgatti and M.G. Everett, "Network Analysis of 2-mode Data," *Social Networks*, vol. 19, no. 3, 1997, pp. 243–269.

Hady W. Lauw is a postdoctoral researcher at Search Labs, Microsoft Research. His research interests include social networks and mining user-generated content to improve search. Lauw has a PhD in computer science from Nanyang Technological University. Contact him at hadylauw@microsoft.com.

John C. Shafer is a senior researcher at Search Labs, Microsoft Research. His research interests include using data mining to improve search quality and developing a next-generation platform for searching over semi-structured data. Shafer has a PhD in computer science from the University of Wisconsin-Madison, with most of his thesis work occurring while he was a member of the Quest Data Mining group at IBM Almaden Research Center. Contact him at jshafer@microsoft.com.

Rakesh Agrawal is a technical fellow and head of Search Labs, Microsoft Research. He's well-known for developing fundamental data mining concepts and technologies and pioneering key concepts in data privacy, including the hippocratic database, sovereign information sharing, and privacy-preserving data mining. Agrawal has a PhD in computer science from the University of Wisconsin-Madison. He's a member of the US National Academy of Engineering and a fellow of the ACM and the IEEE. Contact him at rakesha@microsoft.com.

Alexandros Ntoulas is a researcher at Search Labs, Microsoft Research. His area of expertise is Web information systems, and his research interests include systems and algorithms that facilitate the monitoring, collection, management, mining, and searching of information on the Web. Ntoulas has a PhD in computer science from the University of California, Los Angeles. Contact him at antoulas@microsoft.com.



Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.