World Wide Web (2012) 15:213–232
DOI 10.1007/s11280-011-0143-3

A study of homophily on social media

Halil Bisgin · Nitin Agarwal · Xiaowei Xu

Received: 4 September 2010 / Revised: 26 July 2011 /
Accepted: 4 August 2011 / Published online: 19 August 2011

Abstract The fact that similarity breeds connections, the principle of homophily, has been well-studied in existing sociology literature. Several studies have observed this phenomenon by conducting surveys on human subjects. These studies have concluded that new ties are formed between similar individuals. This phenomenon has been used to explain several socio-psychological concepts such as, segregation, community development, social mobility, etc. However, due to the nature of these studies and limitations because of involvement of human subjects, conclusions from these studies are not easily extensible in online social media. Social media, which is becoming the infinite space for interactions, has exceeded all the expectations in terms of growth, for reasons beyond human mind. New ties are formed in social media in the same way that they emerge in real-world. However, given the differences between real world and online social media, do the same factors that govern the construction of new ties in real world also govern the construction of new ties in social media? In other words, does homophily exist in social media? In this article, we study this extremely significant question. We propose a systematic approach by studying three online social media sites, BlogCatalog, Last.fm, and LiveJournal and report our findings along with some interesting observations. The results indicate that the influence of interest-based homophily is not a very strong leading factor for constructing new ties specifically in the three social media sites with implications to strategic advertising, recommendations, and promoting applications at large.

H. Bisgin
Applied Science Department, University of Arkansas at Little Rock,
Little Rock, AR 72204, USA
e-mail: hxbisgin@ualr.edu

N. Agarwal (✉) · X. Xu
Information Science Department, University of Arkansas at Little Rock,
Little Rock, AR 72204, USA
e-mail: nxagarwal@ualr.edu

X. Xu
e-mail: xwxu@ualr.edu

Keywords homophily · social network · social media · frequent itemset mining · community extraction · modularity · dyadic relations · tags · content analysis · topic modeling · MALLET · NDCG

1 Introduction

Originally proposed by [22], the term *homophily* refers to the concept that similar individuals are assumed to associate with each other more often than others. Since then hundreds of studies have been performed as summarized in [25] that extensively investigated the phenomenon of homophily. Over the years, sociologists have studied the human population on numerous sociodemographic dimensions including race, gender, age, social class, and education and have concluded that friends, co-workers, colleagues, spouses, and other associates tend to be more similar to each other than randomly chosen members of the same population. This phenomenon has been widely used to explain certain sociology concepts like segregation, social mobility, etc.

All these studies have one thing in common, that is all of them were conducted in a physical world scenario by surveying a group of human subjects. Often these subjects belonged to a specific geographical location. These subjects were studied over a set of sociodemographic dimensions as mentioned above. Their ties were subject to social influence. For example, parents had to approve their kids' friends, individuals usually acquainted with those either in the same workplace, schools, etc., that inherently favored the conclusions of the study. Lack of a platform where individuals can explore relations outside their geographical locations, outside their social circles, outside their workplace or schools etc., made it difficult to generalize about the results.

Homophily was categorized by [22] into status homophily and value homophily. Status homophily considered the social status of individuals, implying that individuals with similar social status tend to associate with each other. Value homophily considered similarity based on what people think, implying that individuals who think alike tend to associate with each other regardless of the difference in their social status. However, results from both the categories were based on a small-scale experiment involving human subjects in a physical world scenario.

The advent of social media has offered new strategies to evaluate these existing hypotheses on a much wider scale.Through reactive interfaces, low barrier to publication, and zero operational costs, which are all made possible by the new paradigm of Web 2.0, social media has experienced a phenomenal growth in user participation leading to a participatory web or citizen journalism. Blogosphere, for instance, has been growing at a phenomenal rate of 100% every 5 months [32]. BlogPulse has tracked over 160 million blogs till November 2010 [5]. Facebook recorded more than 500 million active users as of September 2010 [10]; Twitter amassed nearly 200 million users in March 2011 [27]; and other social computing applications like Digg, Delicious, StumbleUpon, Flickr, YouTube, etc., are also growing at a terrific pace. This clearly shows the awareness of social computing applications among individuals.

Often on social media, information such as age, gender, education, and social status is either unavailable or untrustworthy. Moreover, individuals share their interests, likes, dislikes, opinions, perspectives, thoughts, etc. Due to the absence of sociodemographic dimensions, it is difficult to assume homophily that was studied on sociodemographic dimensions. Interests of individuals are one of the strongest factors for evaluating homophily in the virtual world, which was often neglected in the studies conducted in physical world. Precisely because of this reason it is difficult to evaluate status homophily in the virtual world. Authors in [20] studied a university campus social network and concluded that social ties are often influenced by triadic and focal closures.

Another major difference between studies conducted in physical and virtual worlds is the scale of the study. Millions of individuals could be easily studied in the virtual world as compared to the physical world. This makes the results much more conclusive and generalizable. Next we summarize the differences between physical and virtual world in Table 1.

Inspired by the differences between the physical and the online world, in this paper we study the existence of homophily in online social networks. Specifically, we investigate the question of whether construction of ties is inspired by homophily. In other words, we investigate whether individuals are likely to become friends if they share similar interests. We make the following contributions in this work:

- We studied and observed the characteristic differences between factors that influence real world ties and online friendships;
- We developed a methodology to study objectively the impact of these differences in evaluating the factors that affect the ties in online environments. The methodology is generic and can be used to study homophily in any online environment, such as social networking sites. In this regard, we tried to expand our study in some representative platforms that belong to possible categories of social media sites defined by Kaplan and Haenlein [17];
- We proposed some measures to quantify homophily factors along the ties, especially the interest homophily for interests which is the focus of study in our

Table 1 Differences between physical and online/virtual world scenarios.

Physical world	Online/virtual world
Sociodemographic dimensions such as age, gender, education, social status used to study homophily.	Sociodemographic dimensions are often not available or could not be trusted.
Physical locality such as geographical proximity and organizational locality such as workplace, schools play a significant role in governing new ties.	Interactions between individuals span all geographical barriers across different timezones. Geographical or organizational proximity do not govern construction of ties.
User interests, opinions, thoughts, perspectives, and preferences were often ignored in studies conducted in physical world scenario.	Individuals on social media are defined by what they write/share. Interests, opinions, thoughts, perspectives, and preferences are the significant dimensions that could govern new ties.
Construction of new ties in the physical world are often regulated by social status or class.	Construction of ties in virtual world are beyond social status and class.
Studies conducted in the physical world were often limited to a particular geographical area constraining the scale of the study.	Millions of individuals could be easily studied in the virtual world as compared to the physical world. This makes the results much more conclusive and generalizable.

research. However, the measures are also generic and can be easily employed to quantify other forms of homophily, such as geographical, demographic, socio-economic ones;

- We analyzed interest-based homophily at two different levels of granularity - dyadic-level and community-level. Interests were captured using both explicitly mentioned choices and those implicitly derived from the content shared by the individuals using the content analysis approach, probabilistic topic modeling;
- We studied the random rewired network model to compare and contrast the ties-construction process with organically observed ties-construction in online social networks;
- We experimented the proposed methodology rigorously with three online social networks BlogCatalog, Last.fm, and LiveJournal at both dyadic and community level. The organic ties-construction process at these online social networking sites was compared with the random-rewired network model; and
- Finally, results indicated a weak influence of interest-based homophily on the construction of new ties at all three websites. This outcome has the implication that promoting, advertising, and recommendation over the social media should be reconsidered. This is discussed more broadly in Section 8.

The rest of the paper is organized as follows. Section 2 lists the related work in this domain. Section 3 describes the data collection from three online social networks, viz., BlogCatalog, Last.fm, and LiveJournal. Before analyzing complex community affiliations, we performed a preliminary study at a finer granularity to evaluate the hypothesis on online social network data by examining dyadic relations in Section 4. Section 5 examines homophily from a complex perspective of community formation. Specifically, we test the hypothesis that individuals with similar interests are more likely to create ties with each other and that communities emerge when a group of individuals has more links among themselves as compared to the whole population. We study some of the most widely used community extraction algorithms (Fast Modularity and Graclus) and analyze whether the extracted communities actually shared similarities. Next, we compare homophily with a random process from the community formation perspective in Section 6. Section 7 examines homophily along a different dimension of content provided by the users as opposed to the tags in previous experiments. We discuss the possible implications and applications of the findings in Section 8 and we conclude the paper with some future directions in Section 9.

2 Related work

There has been a significant body of work studying the homophily principle using real-world data. This involved conducting surveys with human subjects and then evaluating their responses [16, 22, 31]. As mentioned in the beginning of the paper, often choices for constructing new ties in the real world are influenced by several factors, such as demographics, geographical and organizational locality, etc. There are also some studies confirming that there are also evidences of homophily for those factors like ethnicity, religion, age, and country in online social networks [30].

McPherson and Smith-Lovin discussed the similarity between individuals and ties. Authors studied the various sociodemographic characteristics and the role they play in determining construction of new ties. This article investigated the sources of homophily such as, social structures and cognitive processes. The article also studied the influence from geographical and organizational locality factors. However, the authors did not consider the interests of individuals in governing the ties. The authors also proposed the need for further study to investigate the principle of homophily [25]. Similarly, Singla and Richardson [28] studied the instant messaging data and concluded that friends tend to share similar demographic characteristics. However, interests of these users were not included in the study. In another study, Gilbert and Karahalios consider a set of 35 Facebook users and proposed a regression model for predicting the friendship on Facebook. The features mainly consisted of user demographics and interactions, but did not include their interests [12]. Moreover, the results from a survey of 35 users are not easily extensible when compared to the datasets used in our work.

In the research that Nowak and Rauh conducted, homophily was one of the parameters along with androgyny, anthropomorphism, credibility, and attraction by which the authors tried to determine the influence of the avatar. The experiment did not consider the friendship ties between any two friends. In fact, the purpose was to indicate how a user feels close herself to the avatar to whom she was exposed. What they did was to measure the perception of users by asking them to mark a scale that showed how well those parameters reflect their perception. Finally, the response would determine the level of homophily [26].

Fiore and Donath did a research on an online dating site, the "Site", where they had the profile information of users. Self-reported preferences and the private messaging data were also used in their work. Their analysis pointed out that the users' dating attributes were very similar to the offline world in terms of homophily. In other words, people continued to show very similar tendencies in their search for a mate in the online world [11].

Crandall et al. studied the LiveJournal and Wikipedia data and used activities such as user edits to evaluate the similarity between individuals [8]. This is quite different from the research conducted in this paper which looks as the interests of the users to investigate homophily. In another work by Adamic and Adar, the homepages of users were studied and friendship was modeled using hyperlinks between the homepages [1].

In this paper, we studied online social networks where such factors may not play a significant role. This differentiates our work from the existing works mentioned above in both terms of methods and the data we used from different social media sites. Next, we summarize these works and point out the essential differences.

3 Data collection

Extensibility of any method requires experimenting with it in different domains. Therefore, based on the taxonomy introduced by Kaplan and Haenlein [17], we studied three representative social media sites some of which can be thought as a combination of those possible categories. Among these categories, we used BlogCatalog[1] and LiveJournal[2] as samples for the blogs. As a sample from content

[1] www.blogcatalog.com

[2] http://www.livejournal.com/

communities, Last.fm[3] is another site we utilized. Referring to the taxonomy, those sites can also be considered as social networking sites since they allow users to build friendship ties in the virtual world.

BlogCatalog is a blogging portal where bloggers can submit their blogs, tags, categories, and specify their friends. This data set was obtained from Social Computing Data Repository [33]. The second data set was constructed by crawling Last.fm. Last.fm is a social networking website where users can specify the genre of music they like and connect with others. It hosts a huge community of users and their taste in music. Users specify their friends on Last.fm. This link structure was used to crawl data in a breadth-first fashion. The crawler was forcefully terminated after 279,678 users were crawled. The crawler collected both the network information and the music genre(s) the user likes. While BlogCatalog has a very broad spectrum of interests for users Last.fm has a very narrow focus on user interests. The last data set, LiveJournal, is another virtual platform where the users can keep a blog, journal, or diary. As in the first two social media sites, users can build friendship ties. However, there's no ground truth or systematic way of tagging. Therefore, users are free to define new as many interests as they want. We obtained the friendship network and the interests associated with each user by using the API of the site.

3.1 Data pre-processing

Users in BlogCatalog are required to label their blog(s) by using system-defined tags. There were 344 total tags found in the dataset. Two default tags, *Personal* and *Blogging*, which could be considered as noise were removed. Thus, we accepted the remaining 342 tags as ground truth of tags or interests. Out of 88,784 users on this site, 79,115 users had valid labels for their blogs. In other words, the result of the tag validation process which was a superset of the friendship network has been considered as our initial step.

Last.fm data had a more challenging noise issue when compared to BlogCatalog. Namely, there was no system-defined tagging procedure, which resulted in a high variety of user-defined labels. This required standardization of user-defined tags/genres with respect to a ground truth. Therefore, we assumed a genre reference from Wikipedia where we had 1,496 types of music listed. These music types or genres formed the ground truth list for Last.fm users. In this case, as a superset of a social network to be mined, 64,805 users passed the tag validation stage among the crawled dataset of 279,678 users.

As we stated earlier, LiveJournal had no structured tagging. We were able to filter the invalid tags which do not exist in our ground truth (genre list) for Last.fm, but we were not able to accomplish the same thing for LiveJournal due to the absence of such a ground truth. Therefore, we reserved this data set for the content based analysis in Section 7 which was only applicable to BlogCatalog and LiveJournal.

After discarding the noisy tags (in BlogCatalog) and unrecognized tags (in Last.fm), those users that did not have a single valid tag were removed from the dataset. This resulted in social networks consisting of 78,445 and 54,987 users for

[3] www.last.fm

Table 2 Summary of BlogCatalog and Last.fm networks.

Statistics	BlogCatalog	Last.fm	LiveJournal
Number of nodes	78,445	54,987	50,000
Number of links	1,848,120	214,628	72,234
Link density	0.0006	0.00014	0.00006
Average degree	23.56	3.90	0.69
Attribute name	Category	Genre	NA
Size of attribute domain	342	1,496	NA
Average number of attributes per node	2.49	10.63	NA

BlogCatalog and Last.fm, respectively. The social network datasets obtained from BlogCatalog and Last.fm after pre-processing contain 1,848,120 and 214,628 links respectively. Statistics of all the datasets have been summarized in Table 2.

4 Analyzing dyadic relations

Before analyzing complex community affiliations, we performed a preliminary study at a finer granularity. Specifically, we studied the ties of individuals and analyzed the overlap in the interests of the individuals for both the social networks. We computed the percentage of ties that share common interests and divided the ties into two groups, (a) ties with no common interests and (b) ties with one or more common interests. Table 3 presents the results for both the datasets. Clearly from Table 3, it can be observed that for BlogCatalog data over 84% of the ties do not share a single interest, which indicates that individuals do not consider interests before they form the ties. However, the Last.fm dataset apparently tells a different story. It seems that over 76% of the ties have at least an interest in common. Deeper analysis showed that in Last.fm individuals have an extremely large number and varied forms of interests (as also shown in Table 2, average number of attributes or interests per node is 10.63 for Last.fm as compared to 2.49 for BlogCatalog).

Next, we looked at the normalized similarity score between pairwise individuals connected with a tie by computing Jaccard similarity coefficient, as defined below,

$$J(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{1}$$

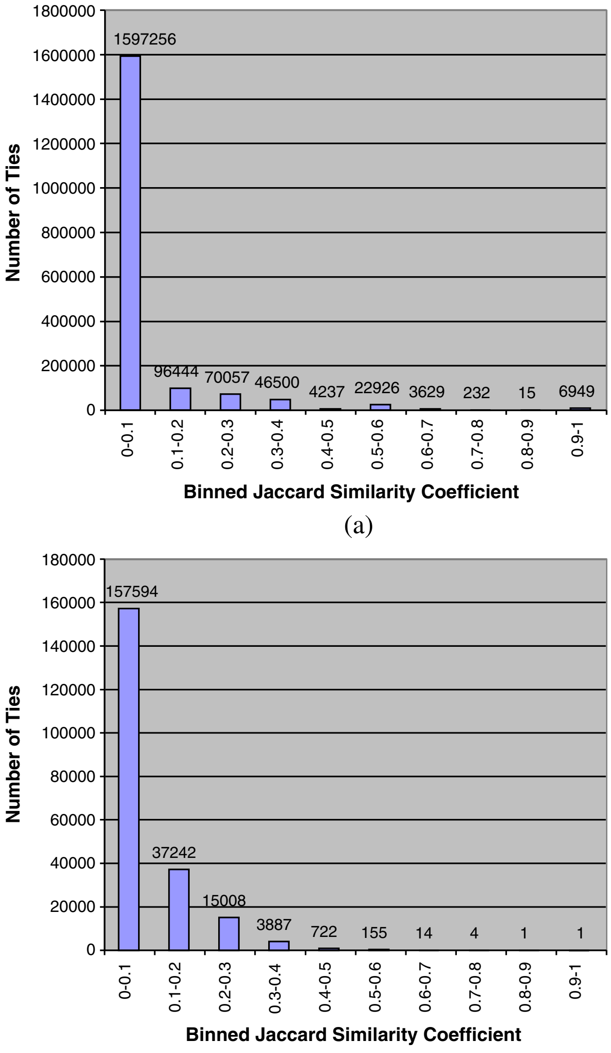
where A and B represents the set of interests of two individuals that share a tie and $0 \leq J(A, B) \leq 1$ is defined as the Jaccard similarity coefficient between the two sets A and B. We then averaged the Jaccard similarity score for all the ties, which was found to be 0.04 for BlogCatalog and 5×10^{-7} for Last.fm. This shows that there is even lesser similarity in terms of interests between individuals who create ties in Last.fm as compared to BlogCatalog. This also confirms that individuals on Last.fm

Table 3 Overlap in interests for individuals sharing ties in BlogCatalog and Last.fm datasets.

	BlogCatalog (%)	Last.fm (%)
0 interests in common	84.0609	23.2490
1 or more interests in common	15.9390	76.7509

have a large number of varied interests which are rarely common between indivudals who share a tie.

We further analyzed the distribution of Jaccard similarity coefficient between the individuals sharing a tie for both the datasets. We binned the similarity scores into equal-sized bins of 0.1 from 0 to 1 and plotted the frequency of ties that fall into the bins. The results for BlogCatalog and Last.fm are shown in Figure 1a and b, respectively. Clearly Figure 1a and b show that over 86% of ties in BlogCatalog dataset and over 73% ties in Last.fm dataset connect individuals with similarity less than 0.1. This study shows that very often individuals that share a tie do not share



(a)

(b)

Figure 1 Distribution of Jaccard similarity coefficient in equal-sized bins of 0.1 for **a** BlogCatalog dataset and **b** Last.fm dataset.

interests, hence contradicting the assumption that homophily influences creation of new ties.

5 Analyzing community structures

Coming together of similar group of people to form communities has been well studied. This has been the underlying phenomenon for the vast literature on community extraction [2, 13, 21, 23, 35]. The micro-level processes of creating new ties based on their similarity gives rise to macro patterns of associations, also known as communities. This concept has been extensively used in discovering communities in online social networks. We studied some of the most widely used community extraction algorithms and analyzed whether the extracted communities actually shared similarities. Next, we briefly describe the community extraction algorithms used in our work, Fast Modularity [7] and Graclus [9].

5.1 Community extraction algorithms

Fast Modularity Unlike other methods, Fast Modularity can extract communities from very large networks due to its hierarchical fashion [7]. It tries to optimize a modularity value during the procedure in an agglomerative way. If we let v and w denote vertices, and A_{vw} represents the the entry in the adjacency matrix with m edges, Clauset et al. defines the modularity function, Q as follows

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \tag{2}$$

where k_v stands for degree of v and c_v represents the cluster that v belongs to. The *Kronecker delta* function, $\delta(i, j)$ is also defined as

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The procedure adopts a bottom-up approach by assuming every node as a community at the beginning and merges the communities if they contribute in Q. Formally, they define the following two quantities which help us represent Q in a more explicit way. The fraction of the edges joining nodes in cluster i and in cluster j is represented by e_{ij} whereas the fraction of ends of edges that are attached to nodes in cluster i is denoted by a_i

$$e_{ij} = \frac{1}{2m} \sum_{vw} A_{vw} \delta(c_v, i) \delta(c_w, j) \tag{3}$$

$$a_i = \frac{1}{2m} \sum_{v} k_v \delta(c_v, i) \tag{4}$$

Clearly, $\delta(c_v, c_w)$ can be expressed as $\sum_i \delta(c_v, i)\delta(c_w, i)$. Hence, (2) can be rewritten as,

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\sum_i \delta(c_v, i)\delta(c_w, i) \\ &= \sum_i \left[\frac{1}{2m}\sum_{vw} A_{vw}\delta(c_v, i)\delta(c_w, i)\right] \\ &\quad - \sum_i \left[\frac{1}{2m}\sum_v k_v\delta(c_v, i)\frac{1}{2m}\sum_w k_w\delta(c_w, i)\right] \\ &= \sum_i (e_{ii} - a_i^2) \end{aligned} \tag{5}$$

As the cluster configuration is changed due to agglomeration Q is recomputed. Configuration corresponding to the maximum value of Q is selected as the best partitioning result.

Graclus Spectral clustering has been a well-studied partitioning approach for graphs. However, its computational cost has led people to improve this approach. Graclus algorithm is a result of those studies where Dhillon et al. tried to model their problem in an equivalent form [9]. In particular, instead of working with an eigenvalue based approach, they have utilized a *k-means* approach. More specifically, Graclus is based on a *kernel k-means* clustering whose performance has been shown to be much better than spectral clustering methods.

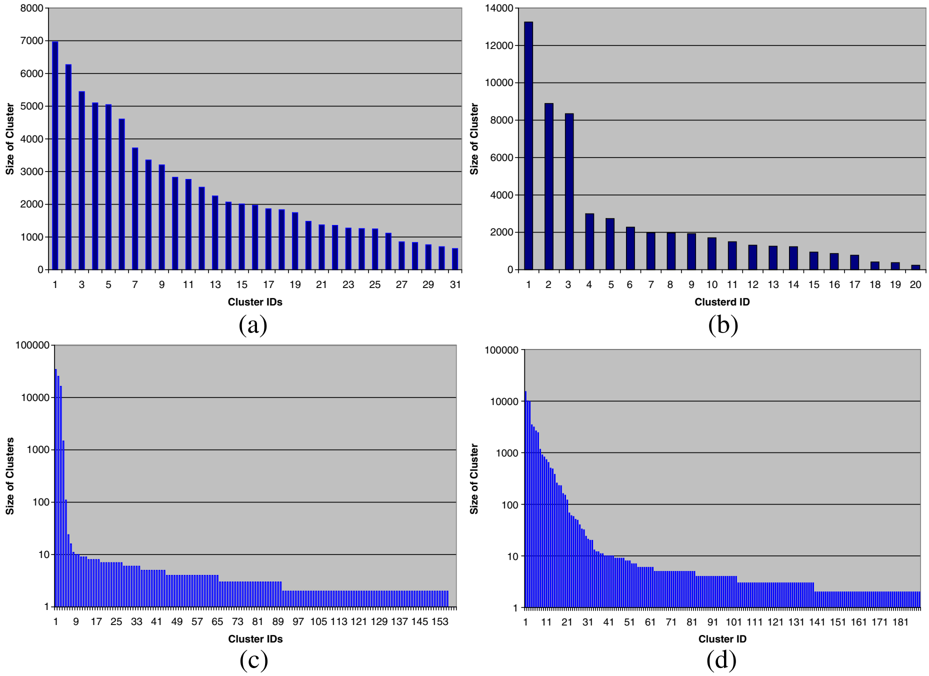
It outperforms not only in terms of time but also in terms of memory and quality. Its advantage is considerable compared to the spectral fashion where for k eigenvectors and n data points $O(nk)$ storage is needed.

Another characteristics of Graclus algorithm is its multilevel fashion. In other words, it initially performs a clustering on a coarse graph and refines it in the refinement stage. In contrast to previous multilevel algorithms Metis [18] and Chaco [14], which are based on the goal of optimization of Kernighan–Lin objective [19], Graclus is not constrained to equal-sized clusters.

5.2 Methodology

In this section, we present the experimental methodology to study homophily with respect to community formation. We extracted the communities from the social network datasets and investigated whether creation of these ties was influenced by the similarity of interest(s). Thus, we first identify the communities and then extract the interests of these communities.

Community structure detection We applied two clustering algorithms, viz., Graclus and Fast Modularity, to obtain communities for each social network dataset. Graclus extracts communities using a multilevel approach, whereas Fast Modularity uses a completely different approach of splitting the network as explained in Section 2. Graclus requires the total number of clusters a priori, whereas Fast Modularity automatically computes the number of clusters. Graclus tries to partition the data into equal-sized clusters, whereas Fast Modularity could partition the data into highly



(a) (b) (c) (d)

Figure 2 Distribution of cluster sizes obtained using **a** Graclus on BlogCatalog dataset, **b** Graclus on Last.fm dataset, **c** Fast Modularity on BlogCatalog dataset, and **d** Fast Modularity on Last.fm dataset.

uneven cluster distribution. After careful analysis of data, it was found that the optimal number of clusters for BlogCatalag and Last.fm was 31 and 20, respectively. Cluster distribution obtained from both the clustering methods on both datasets are illustrated in Figure 2a–d. One can see that Fast Modularity generates a large number of highly uneven-sized clusters with several clusters sized less than 100, whereas clusters obtained through Graclus are comparatively fewer and sufficiently large.

Shared interests acquisition Next we compare the extracted community to the whole population or the entire dataset with respect to interests as specified by the individuals using categories and tags. To extract representative interests of communities as well as the entire population, we utilized a frequent pattern mining algorithm. An apriori algorithm [6] was implemented to find out frequently occurring itemsets for each cluster (also the community), as well as the entire population from each dataset

Table 4 Number of clusters for BlogCatalog and Last.fm for Graclus and Fast Modularity considered for interests acquisition.

	BlogCatalog	Last.fm
Graclus	31	20
Fast Modularity	5	20

Table 5 Top ten tags from BlogCatalog and clusters found by Fast Modularity and Graclus.

BlogCatalog	Fast Modularity	Graclus
Entertainment	Sports	Technology
Art	Travel	Internet
Technology	Technology	Lifestyle
Internet	Entertainment	Art
Marketing	Real Estate	Entertainment
Health	Finance	Writing
Travel	Music	Marketing
Music	Business	Music
Lifestyle	Art	Photography
Business	Internet	Family

with a minimum support of 1% of the cluster size. For this step we considered the clusters mentioned in Table 4.

5.3 Experimental results

This section presents the results of the community aspect of the homophily phenomenon. We compute the overlap between the interests of the communities and the entire population for all the datasets. We analyze the results using qualitative and quantitative metrics. The most common interests shared within a community and the ones that are more prominent in the population need to be compared according to the previous steps. Therefore, we first fixed the frequent items of populations as ground truth with respect to their frequencies. The apriori algorithm showed that there are 79 tags in BlogCatalog and 189 in Last.fm satisfied the minimum support, and we assumed these two ordered sets as ground truth whose top ten tags are showed in Tables 5 and 6.

The processed 76 groups also gave tags in an order which is specific to that community. Whereas in some clusters there are more crowded lists, we notice fewer tags in the others. Similarly, groups may contain common interests, but in different orders.

A valid measure which can quantitatively prove how communities are similar to others is the *discounted cumulative gain (DCG)* method. As previously mentioned, regarding the ground truth relevance scores are assigned to every five labels in the list. Depending on the list sizes of two social media sites, scorings start from 36

Table 6 Top ten tags from Last.fm and clusters found by Fast Modularity and Graclus.

Last.fm	Fast Modularity	Graclus
Rock music	Rock music	Pop music
Electronic music	Indie music	Dance music
Indie music	Electronic music	Rock music
Experimental music	Folk music	Soul music
Pop music	Pop music	Electronic music
Folk music	Indie rock	Indie music
Jazz	Experimental music	Pop rock
Alternative rock	Electronica	Hip hop music
Electronica	Alternative rock	Jazz
Indie rock	Classic rock	Alternative rock

Table 7 Statistics.

	BlogCatalog		Last.fm	
	μ	σ	μ	σ
Graclus	0.95	0.05	0.98	0.01
Fast Modularity	0.96	0.04	0.98	0.008

and 16 for BlogCatalog and Last.fm, respectively. What *DCG* does is to calculate information gain by the following expression,

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \qquad (6)$$

where i denotes the rank position and rel_i represents the relevance score of the specific item.

Since the ground truth lists have the maximum information gain, we call value obtained from those as *ideally discounted cumulative gain*, *IDCG*. If one takes the ratio of *DCG* and *IDCG*, outcome will be the normalized *DCG* and denoted by *nDCG*.

$$nDCG = \frac{DCG}{IDCG} \qquad (7)$$

We took processed 76 clusters into account to calculate *nDCG* in order to measure the similarity between any group and the general site tendency. Since the label list may differ from group to group, we used the overlapping tags through all calculations preserving the rankings and relevance scores. Although there exist tags that appear in partitions, but not in the bigger set and vice versa, the amount of overlapping patterns led to a reliable comparison. As a matter of fact, our results show that any group of people constituting a community have a very high similarity with the population they come from. Table 7 summarizes similarity values acquired from both clustering algorithm results for both social media sites.

6 Random Rewired analysis

Both the community extraction algorithms, viz., Graclus and FastModularity, recognize a group of individuals as a community if they have more links among them as compared to the whole population. Now results from the previous section show that different communities have highly similar interests. This indicates that individuals connect with each other regardless of their interest(s). To further examine this, we performed another study. We broke all the ties that were created by the individuals in both the datasets, viz., BlogCatalog and Last.fm, and created new ties following the BA model [3]. The Network Workbench tool was used to create the random rewiring [29]. The new random networks were constrained to the same set of nodes and approximately the same number of ties correspondingly in the two original datasets. The network characteristics for the random rewired dataset for BlogCatalog and Last.fm are shown in Table 8.

We followed the same experimental methodology explained in Section 5.2. For both the random rewired datasets we ran Graclus and FastModularity to identify communities. Note that clustering parameters were kept exactly the same as with

Table 8 Summary of Random Rewired models of BlogCatalog and Last.fm networks.

Statistics	BlogCatalog	Last.fm
Number of nodes	78,445	54,987
Number of links	1,878,849	214,628
Link density	0.00061	0.00014
Average degree	23.95	3.90

the original dataset in Section 5.2. The cluster distribution for the random rewired datasets for BlogCatalog and Last.fm using both Graclus and FastModularity are shown in Figure 3a–d. The cluster distribution for random rewired dataset is very similar to the original dataset as shown in Figure 2a–d. We observed a very similar power law distribution, where there are several clusters with very few cluster members and very few clusters with a large number of cluster members. Clusters with less than 100 members were ignored for further analysis, as in Section 5.2. Table 9 shows the clusters that were considered for the interests acquisition step. Once the communities were extracted, we identified the communities' interests using the approach mentioned in Section 5.2 and computed the overlap between communities' interests and interests of the whole population for both BlogCatalog and Last.fm. We analyzed the top-k interests of individuals in a community and the entire population using tag clouds for varying k. Due to the space constraints, we display some sample tag clouds obtained using Fast Modularity clustering approach for the random rewired datasets for BlogCatalog and Last.fm in Figure 4 for $k = 50$.

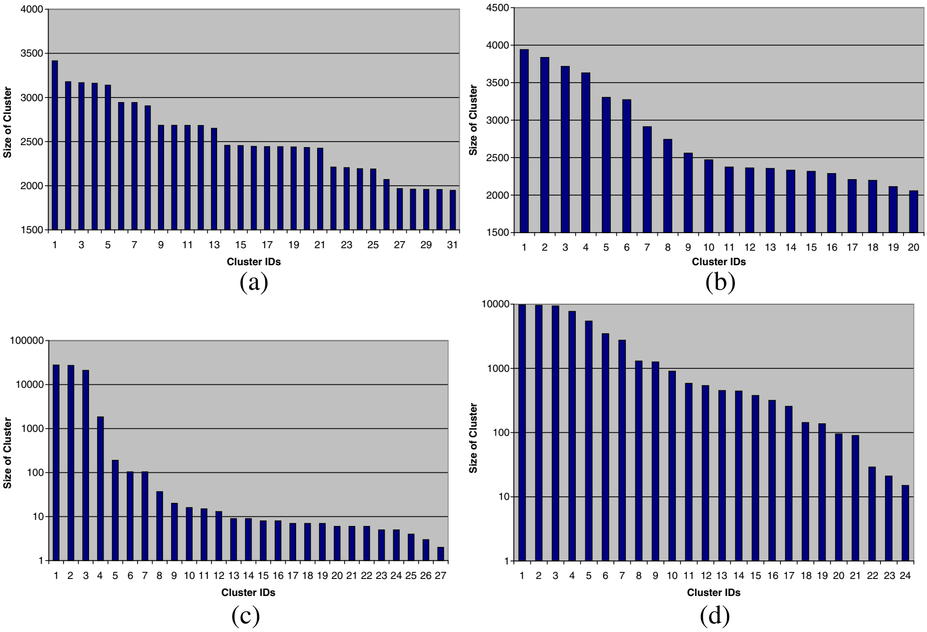


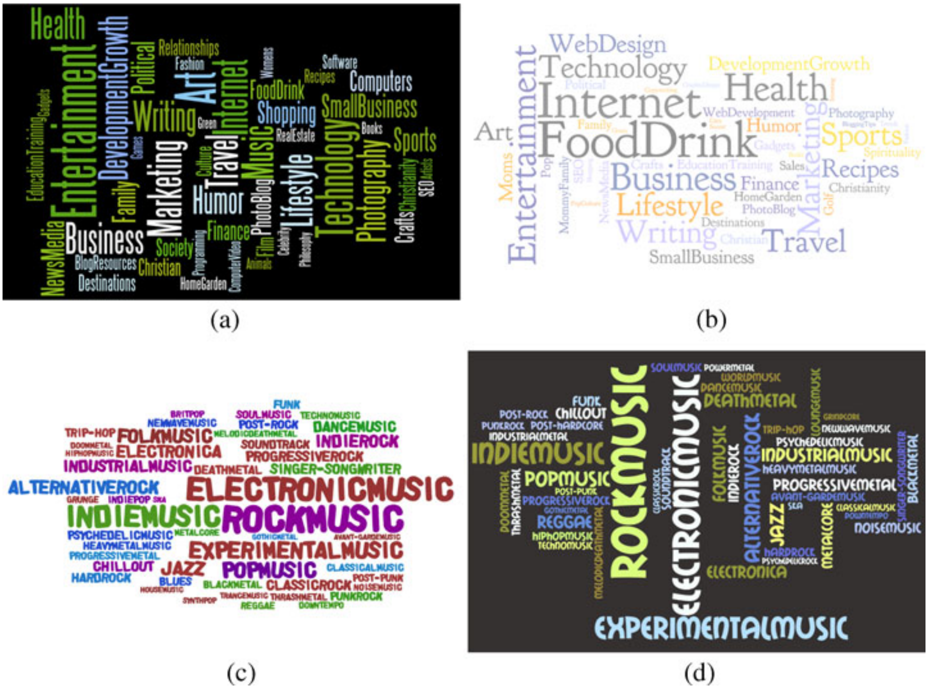
Figure 3 Distribution of cluster sizes obtained using **a** Graclus on Random Rewired model of BlogCatalog dataset, **b** Graclus on Random Rewired model of Last.fm dataset, **c** Fast Modularity on Random Rewired model of BlogCatalog dataset, and **d** Fast Modularity on Random Rewired model of Last.fm dataset.

Table 9 Number of clusters for Random Rewired models of BlogCatalog and Last.fm for Graclus and Fast Modularity considered for interests acquisition.

	BlogCatalog	Last.fm
Graclus	31	20
Fast Modularity	7	19

It can be also inferred from the figures that interest rankings for the randomly rewired network are not very different from the original rankings obtained before.

Due to the visualization limitations of tag clouds, we compared the interests of different communities with the whole population using the NDCG measure as explained in Section 5.3. We report the mean NDCG values and the variance results in Table 10. It can be observed from the NDCG values that there is a high similarity in interests between different communities and the whole population. These results are quite similar to what we obtained from original BlogCatalog and Last.fm datasets where the creation of ties was not random. This clearly demonstrates that the creation of ties in online communities does not depend on the interest(s), but is rather a random process. This research has raised more questions than it has answered. We need to investigate further into the factors that influence the creation of new ties in online communities, which is discussed more in Section 9 as a future research direction.



(a) (b)

(c) (d)

Figure 4 Top 50 tags from BlogCatalog and Last.fm clusters **a** largest cluster from BlogCatalog, **b** smallest cluster from BlogCatalog, **c** largest cluster from Last.fm, **d** smallest cluster from Last.fm.

Table 10 *NDCG* statistics comparing communities with the whole population for Random Rewired model of BlogCatalog and Last.fm networks.

	BlogCatalog		Last.fm	
	μ	σ	μ	σ
Graclus	0.996	0.001	0.945	0.013
Fast Modularity	0.981	0.023	0.918	0.033

7 Content based analysis

So far we have investigated homophily in BlogCatalog and Last.fm in which it was possible to identify groups based on community extraction algorithms and measure the similarities using the tags as feature vectors. Another way to look at the dyadic relation is to compare users' interests based on the content that they share with others. Between those two data sets, this comparison method was only applicable to BlogCatalog, where we have users' written expressions. The other dataset from Last.fm used in our previous experiments does not allow users to share content. So for a more generic analysis, we seek out another online social network website, LiveJournal, in addition to BlogCatalog. The data from LiveJournal is collected using their API service. The LiveJournal data consisted of 50,000 users and 72,234 friendship ties. To refresh, the BlogCatalog data consisted of 78,445 users and 1,848,120 friendship ties (see Table 2).

In our previous experiments in Section 4, each user was represented by a feature vector that was built using their tags. In this study, we used the users' blog post text to construct feature vectors. Every user's blog post text was treated as a document. Multiple blog posts for a user were concatenated to create a single document, leading to one document per user. We also generated documents for each LiveJournal user which were consisting of their interests. This resulted in 78,445 documents for BlogCatalog users and 50,000 documents for LiveJournal users. Probabilistic topic modeling technique was used to construct feature vectors for the users. We explain this technique followed by the findings from dyadic analysis of these content-based feature vectors.

7.1 Feature extraction by topic modeling

In text mining domain, different methodologies including bag of words, Latent Dirichlet Allocation, and latent semantic analysis have been used to construct feature vectors for supervised and unsupervised machine learning tasks. Whereas bag of words is a naïve approach, the latter approaches are more sophisticated leveraging probabilistic topic modeling techniques.

If $P(z)$ is the distribution of topics and $P(w|z)$ is to be the distribution of the words given the topic z, then we can write the probability that ith word, w_i is sampled from the topic $z_i = t$ with $P(w_i|z_i = t)$. When all the topics are taken into account, the following expression becomes the distribution of words within a document

$$P(w_i) = \Sigma_{t=1}^{T} P(w_i|z_i = t)P(z_i = j) \tag{8}$$

where T represents the number of topics.

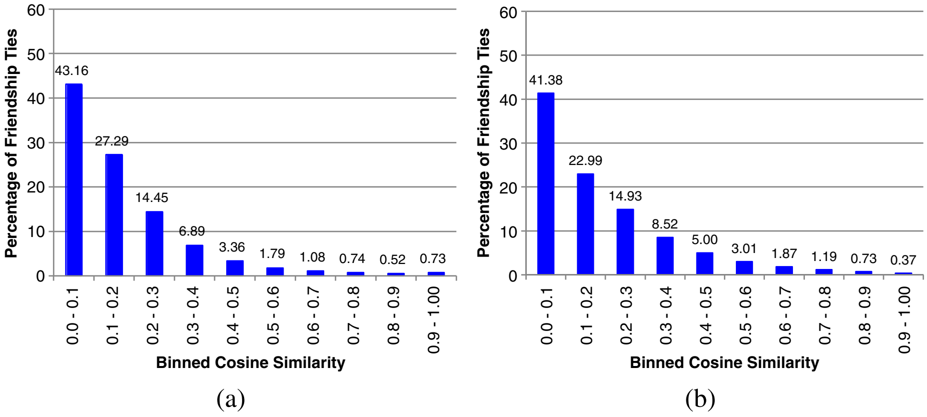
This model mainly tries to find each multiplicative term in the summation. The first term indicates words that are important for which topic and the second one gives the information on the topic distribution over documents. For document d, $\theta(d) = P(z)$ stands for the multinomial distribution over topics. In the model called Probabilistic Latent Semantic Indexing method (pLSI) there is no assumption about how the θs are generated [15]. In another model called Latent Dirichlet Allocation (LDA) by Blei et al. [4], there is a Dirichlet prior on θ that makes not only the inference step more convenient, but also the model more generalized to new documents.

In this part of the study, we have used LDA to obtain the parameter θ for every document, which is the blog content in our case. Then we represented it for each user as a probability vector on T topics. We used MALLET [24] for accomplishing this task. MALLET is a software package used in many machine learning and text mining applications. In order to convert the text to features, we preferred using the topic modeling feature of the package. MALLET was fed by the documents that we had generated for every user. Hence, as mentioned above, we have 78,445 documents for BlogCatalog users and 50,000 documents for LiveJournal users.

We preferred the default number of topics, which was 100, for our feature vector. In this fashion, we acquired the topic distributions for each document which became our attribute vector for computing the similarity in the next step.

7.2 Results and analysis

As long as we have the same number and same type of features in any data set, we can measure the distance between any two instances using any kind of similarity. Since we already have a feature vector for every user, we computed cosine similarities for each friendship tie. The number of friendship ties are binned in intervals of 0.1 cosine similarity values, and percentages of the friendship ties are computed. The results are shown in Figure 5a and b for BlogCatalog and LiveJournal data, respectively. For both the data sets, over 40% of the friendship ties were only 1% similar in their



(a) (b)

Figure 5 Dyadic analysis using content based features for **a** BlogCatalog data and **b** LiveJournal data.

content, and over 95% of the friendship ties were less than 50% similar. These results are consistent with our previous findings and indicate that content similarity does not breed connections. In other words, friendship ties are not created based on similarity of interests (gauged using content provided by the users).

8 Implications and applications

The findings indicate that being friends does not necessarily require that any two friends have to have similar interests. In other words, connections may not be constructed based on common interests. This conclusion has several implications and applications. First of all, this research helps us in understanding or rather rethinking the principles of segregation or community formation in virtual environments. The outcomes enable us to examine the conventional theories that were established based on the observations limited by physical or real-world settings and improve them or even develop new ones for virtual worlds. This could have consequences in applications such as search and recommendations. Since interests may not be the major motivation behind creation of new ties, perhaps online social networks could incorporate diversity in interests before recommending new friends. Not only recommendation for a friend, but also recommendation for a product within a friendship network may not necessarily need to assume that they have a similar view on the product. In a very recent study by [34], the challenging nature of the recommendation systems was highlighted and a new method to improve these models was also introduced. Besides implications to the fundamental assumptions of the search and recommender systems, the outcome of our study can also be considered as an additional factor.

Further, similarity computation between individuals could include dimensions beyond just the interests, which could have broader impacts in search and information-retrieval applications. Companies can use this idea to design efficient marketing strategies for social media and virtual environments in general. In real word, companies target users and their contacts to advertise products and service assuming that users who are connected share similar interests. However, the increasing participation in online social media and the findings of this work indicates a need to reevaluate advertising and marketing strategies. In the virtual world companies need to be careful in marketing products to a particular group. It is possible that the members of these groups could be close friends but still come from diverse backgrounds and possess different interests.

9 Conclusion

In this paper, we studied the principle of homophily in the context of online social media where interests are the primary data to quantify it. Besides the interests of the users, their blog posts, if applicable, were also taken into account to investigate the homophily. Although the existence of homophily has been verified in the real-world, its existence in online social media is questionable due to the differences pointed out in our work. We proposed a systematic approach to study three online social media networks—BlogCatalog, Last.fm, and LiveJournal—leveraging four independent methodologies, -dyadic, community-based, random rewired, and content-based analysis. Analysis based on dyadic relations demonstrated that the users that share a tie in online social network often do not share interests. For community structure analysis, we specifically extracted communities using two most widely used community discovery algorithms based on the network ties. The emerging communities had very similar interests not only to each other but also to the whole population. This implies that the communities that are evolved based on dense emergence of ties within a specific group of individuals do not have distinctive interests, indicating that the ties that are constructed are not governed primarily by interest homophily. The interests were identified using tags and the content, depending upon availability. For the ones with the content, we utilized probabilistic topic modeling whose results also agreed with the other findings. Further, a random rewired network for these online social networks showed very similar analysis, suggesting towards a random process behind the creation of new ties.

This study has raised several interesting questions, such as, what are the primary factors behind construction of new ties in online social media? Are the real-world ties also influenced by the online social media due to the inevitable penetration of social media in our day-to-day lives? What are the evolutionary characteristics of these ties that warrant a longitudinal study?, among other interesting questions.

Acknowledgements This research was funded in part by the National Science Foundation's Social-Computational Systems (SoCS) Program within the Directorate for Computer and Information Science and Engineering's Division of Information and Intelligent Systems (Award numbers: IIS-1110868 and IIS-1110649) and by the U.S. Office of Naval Research (Grant number: N000141010091). We gratefully acknowledge this support.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)
2. Agarwal, N., Liu, H.: Modeling and data mining in blogosphere. In: Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 1. Morgan & Claypool Publishers
3. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. BlogPulse Stats: http://www.blogpulse.com/. Online; accessed 30 April 2011
6. Bodon, F.: A fast apriori implementation. In: Goethals, B., Zaki, M.J. (eds.) Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), CEUR Workshop Proceedings, vol. 90, Melbourne, Florida, USA, p. 19 (2003)
7. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 66111 (2004)
8. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 160–168. ACM, New York (2008)
9. Dhillon, I.S., Guan, Y., Kulis, B.: Weighted graph cuts without eigenvectors: a multilevel approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1944–1957 (2007)
10. Facebook Statistics—Press Room: http://www.facebook.com/press/info.php?statistics/ (2008). Online; accessed 30 April 2011
11. Fiore, A.T., Donath, J.S.: Homophily in online dating: when do you like someone like yourself? In: CHI'05 Extended Abstracts on Human Factors in Computing Systems, pp. 1371–1374. ACM, New York (2005)

12. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 211–220. ACM, New York (2009)
13. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821 (2002)
14. Hendrickson, B., Leland, R.W.: A multi-level algorithm for partitioning graphs. In: Supercomputing (1995)
15. Hofmann, T.: Probabilistic latent semantic indexing. In: Gey, F., Hearst, M., Tong, R. (eds.) Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), 15–19 August 1999, Berkeley, CA, USA, pp. 50–57. ACM Press, New York (1999)
16. Hoyt, D.R., Babchuk, N.: Adult kinship networks: the selective formation of intimate ties with kin. Soc. Forces **62**(1), 84–101 (1983)
17. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. Bus. Horiz. **53**(1), 59–68 (2010)
18. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. **20**(1), 359–392 (1998)
19. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. **49**(1), 291–307 (1970)
20. Kossinets, G., Watts, D.J.: Origins of homophily in an evolving social network. Am. J. Sociol. **115**(2), 405–450 (2009)
21. Kumar, R., Novak, J., xRaghavan, J., Tomkins, A.: On the bursty evolution of blogspace. World Wide Web **8**(2), 159–178 (2005)
22. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: a substantive and methodological analysis. Freedom and Control in Modern Society **18**, 66 (1954)
23. Lin, Y.R., Sundaram, H., Chi, Y., Tatemura, J., Tseng, B.L.: Blog community discovery and evolution based on mutual awareness expansion. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 48–56. IEEE Computer Society, Piscataway (2007)
24. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
25. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**(1), 415–444 (2001)
26. Nowak, K.L., Rauh, C.: The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. J. Comput.-Mediat. Commun. **11**(1), 153–178 (2005)
27. Shiels, M.: Twitter Co-Founder Jack Dorsey Rejoins Company. http://www.bbc.co.uk/news/business-12889048/ (2011). Online; accessed 14 May 2011
28. Singla, P., Richardson, M.: Yes, There is a Correlation: From Social Networks to Personal Behavior on the Web (2008)
29. Team, NWB: Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan. http://nwb.slis.indiana.edu/ (2006)
30. Thelwall, M.: Homophily in myspace. J. Am. Soc. Inf. Sci. Technol. **60**(2), 219–231 (2009)
31. Verbrugge, L.M.: The structure of adult friendship choices. Soc. Forces **56**(2), 576–597 (1977)
32. White, D., Winn, P.: State of the Blogosphere. http://technorati.com/blogging/feature/state-of-the-blogosphere-2008/ (2008). Online; accessed 30 April 2011
33. Zafarani, R., Liu, H.: Social Computing Data Repository at ASU. http://socialcomputing.asu.edu/ (2009)
34. Zhang, R., Tran, T., Mao, Y.: Opinion helpfulness prediction in the presence of "words of few mouths". World Wide Web, 1–22 (2011). doi:10.1007/s11280-011-0127-3
35. Zhou, Y., Davis, J.: Community discovery and analysis in blogspace. In: Proceedings of the 15th International Conference on World Wide Web, p. 1018. ACM, New York (2006)