Investigating Homophily in Online Social Networks

Halil Bisgin
Applied Science Department
University of Arkansas at Little Rock
Little Rock, Arkansas 72204
hxbisgin@ualr.edu

Nitin Agarwal and Xiaowei Xu
Information Science Department
University of Arkansas at Little Rock
Little Rock, Arkansas 72204
{nxagarwal, xwxu}@ualr.edu

Abstract—**Similarity breeds connections, the principle of homophily, has been well studied in existing sociology literature. 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. New ties are formed in social media just like the way 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 two online social media sites, BlogCatalog and Last.fm and report our findings along with some interesting observations.**

I. Introduction

Homophily [12] states that similar individuals associate with each other more often than others. Several studies have been performed since as summarized in [14] 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 concluded that friends, coworkers, colleagues, spouses, and other associations 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 subject 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 the results.

Advent of social media has offered new strategies to evaluate these existing hypothesis on a much wider scale. Ease of use and low barriers to publications has attracted masses to participate and contribute to social media generating a humongous source of human interactions and intelligently crafted data. People connect with each other beyond geographical barriers, across different timezones diminishing the constraints of physical boundaries in creating new ties. One of the strongest factors leading to homophily in physical world is locality due to geographic proximity. This is one of the major differences between ties in physical world and online/virtual world. Often on social media, information such as age, gender, education, social status is either unavailable or untrustworthy. However, individuals express 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 to evaluate homophily in virtual world which was often neglected in the studies conducted in physical world. 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 virtual world as compared to physical world. This makes the results much more conclusive and generalizable.

Inspired by the differences between physical and online world, in this paper we study the existence of homophily in online social networks. In other words, we investigate whether individuals are likely to become friends if they share similar interests. We propose a systematic study using online social networks and analyze the factors that govern the construction of new ties. Next we describe the community structure algorithms used to identify communities in social networks based on their ties. These communities are then explored to test the hypothesis.

II. Community Structures

It has been well studied that similar group of people come together to form communities. This has been the underlying phenomenon for the vast literature on community extraction [2], [13], [18], [8], [11]. The micro-level processes of creating new ties based on their similarity

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.61

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 study some of the most widely used community extraction algorithms and analyze whether the extracted communities actually shared similarities. Next we briefly describe the community extraction algorithms used in our work, Fast Modularity [4] and Graclus [6].

A. *Fast Modularity*

Unlike other methods, Fast Modularity can extract communities from very large networks due to its hierarchical fashion [4]. It tries to optimize a modularity value during the procedure in an agglomerative way. Let v and w denote vertices, and A_{vw} represents an entry in the adjacency matrix with m edges, [4] defines the modularity function, Q as,

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

where k_v stands for degree of v and c_v represents the cluster that v belongs to. Q is recomputed as the cluster configuration changes due to agglomeration. Cluster configuration with maximum value of Q is selected.

B. *Graclus*

Spectral clustering has been a well studied graph partitioning algorithm. However due to high computational costs researchers have proposed variations. Graclus [6] is one such algorithm where instead of an eigenvalue based approach, authors utilized a *k-means* approach. Graclus is based on a *kernel k-means* clustering that is known to perform much better than spectral clustering methods in terms of time, memory, and quality. Graclus works in multilevel fashion i.e., it initially performs a clustering on a coarse graph and refines it in the refinement stage. Graclus is also not constrained to equal-sized clusters.

III. Data Collection

Two online social networks were used to perform the analysis, BlogCatalog (www.blogcatalog.com) and Last.fm (www.last.fm). BlogCatalog is a blogging portal where bloggers can submit their blogs, tags, categories, and specify their friends. This data was obtained from Social Computing Data Repository [17]. 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. Crawler collected both the network information and the music genre(s) the user likes. While BlogCatalog has a very broad spectrum of interests a user could have, on the other hand Last.fm has a very narrow focus on user interests. Due to user-defined tags, Last.fm has huge set of tags that required standardization. Wikipedia's genre reference was used to discard unrecognized tags. Those users that did not have a single valid tag were removed from the dataset. Statistics of both the datasets have been summarized in Table I.

Table I
Summary of BlogCatalog and Last.fm Networks

Statistics	***BlogCatalog***	***Last.fm***
Number of Nodes	78,445	54,987
Number of Links	1,848,245	214,628
Link Density	0.0006	0.00014
Average Degree	23.56	3.90
Attribute Name	*Category*	*Genre*
Size of Attribute Domain	342	1496
Average number of attributes per node	2.49	10.63

IV. Methodology

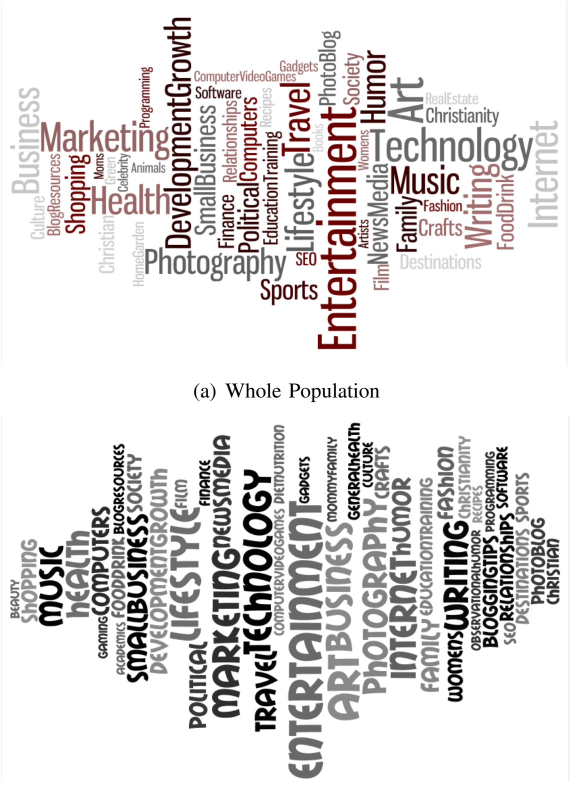
In this section we present the experiment methodology to test the hypothesis that individuals with similar interests are more likely to create ties with each other. It has been widely studied that communities emerge when a group individuals have more links amongst themselves as compared to the whole population. We leverage this phenomenon to extract the communities from the social network datasets and investigate whether creation of these ties were influenced by the similarity of interest(s). Towards this direction, we first identify the communities and then extract the interests of these communities.

A. *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 uneven cluster distribution. It can be observed Fast Modularity generates several clusters with size less than 100, whereas clusters obtained through Graclus are sufficiently large. For further analysis we ignore such clusters with size less than 100.

B. *Shared Interests Acquisition*

Next we compare the extracted community to the whole population or the entire dataset with respect to the interests. To extract the interests of communities as well as the entire population we utilize frequent pattern mining technique. An apriori algorithm [3] was implemented to find out the one



(a) Whole Population

(b) Biggest Cluster

Figure 1. BlogCatalog top 50 tags

item itemsets for each cluster (also the community) as well as whole population from each dataset with a minimum support of 1% of the cluster size.

V. Experimental Results

Here we compute the overlap between the interests of the communities and the entire population for both the datasets, viz., BlogCatalog and Last.fm. We analyze the interests by visualizing their top interests. We analyzed the top-k interests of individuals in a community and the entire population using tag clouds for varying k. Due to space constraints, we only display the tag cloud for the biggest community obtained using FastModularity clustering approach for BlogCatalog in Figure 1. It can be observed from these tag clouds that the interests of a community is not very different from the whole population. Interested readers can found more tag clouds at www.ualr.edu/hxbisgin/Homophily/WI10/Images/

Because of the visualization limitations of tag clouds, we explore other statistical measures such as *normalized discounted cumulative gain (NDCG)* [10]. We report the average $NDCG$ values and the variance in Table II. Our results show that any group of people constituting a community have a very high similarity with the population in terms of interest(s). Such a high similarity value of communities with respect to the entire population imply that the within community similarity is also extremely high. This means that different communities have very similar interests.

Table II
$NDCG$ Statistics comparing Communities with the whole Population for BlogCatalog and Last.fm.

	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

If homophily had an influence over the creation of new ties then different communities would have interests specific to them and significantly different from the other comunities. Given the fact that different communities have similar interests, this contradicts the assumption that homophily influences creation of new ties.

VI. Analyzing Dyadic Relations

For further examination, we performed another study, at a much finer granular level of the online social network data. We analyzed the dyadic relations of individuals.

We computed the normalized similarity score between pairwise individuals sharing a tie by computing Jaccard similarity coefficient. Average Jaccard similarity score for all the ties 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 have a varied and large number of interests which are rarely common between indivudals who share a tie. We further analyze by studying 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. It was found that over 86% of ties in Blogcatalog dataset and over 73% ties in Last.fm dataset connect individuals with similarity < 0.1.

VII. 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 [9], [12], [16]. Often choices for constructing new ties in real-world are influenced by several factors, such as demographics, geographical and organizational locality, etc. In this paper, we studied online social network where such factors do not play much role. This differentiates our work from the existing works mentioned above. However, to the best of our knowledge, there have been very few studies that involved analysis of large online social networks to investigate the principle of homophily.

Authors in [14] study the various sociodemographic characteristics such as social structures and cognitive processes and the role they play in determining construction of new ties. The article also studied the influence from geographical and organizational locality factors. However, authors did not consider the interests of individuals in governing the ties. Similarly, [15] study 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 [7], authors 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. Moreover, the results from a survey of 35 users are not easily extensible, when compared to the datasets used in our work. Authors in [5] study the LiveJournal and Wikipedia data and used activities such as user edits to evaluate the similarity between individuals. 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 [1], authors study the homepages of users and model friendship using hyperlinks between the homepages. While one can link to webpages of several individuals, this does not make the person friends with all of them. Moreover, creating a link does not tell anything about the interests.

VIII. Conclusions and Future Work

In this paper, we proposed a systematic approach to study homophily concept on two online social media networks, BlogCatalog and Last.fm. We extracted communities based on the network ties. The emerged 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 by homophily. This result is also highlighted when the dyadic relations are studied. We plan to expand our study beyond the two online communities. Such data would help in generalizing the conclusion and analyze the problem from different perspectives, identifying various factors that influence construction of new ties and their longevity in virtual and/or physical world. This will also enable us to study the causality relationship between virtual and physical world ties.

References

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] N. Agarwal and H. Liu. *Modeling and Data Mining in Blogosphere.* Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2009.

[3] F. Bodon. A fast apriori implementation. In *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, 2003.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):66111, December 2004.

[5] D. Crandall et. al. Feedback effects between similarity and social influence in online communities. In *ACM International conference on KDD*, pages 160–168. ACM, 2008.

[6] I. Dhillon et. al. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1944–1957, 2007.

[7] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *International conference on Human factors in computing systems*, pages 211–220, 2009.

[8] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[9] D. Hoyt and N. Babchuk. Adult kinship networks: The selective formation of intimate ties with kin. *Social forces*, 62(1):84–101, 1983.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):446, 2002.

[11] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[12] P. Lazarsfeld and R. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18:66, 1954.

[13] Y. Lin et. al. Blog community discovery and evolution based on mutual awareness expansion. In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*, pages 48–56, 2007.

[14] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[15] P. Singla and M. Richardson. Yes, there is a correlation:-from social networks to personal behavior on the web. 2008.

[16] L. Verbrugge. The structure of adult friendship choices. *Social Forces*, 56(2):576–597, 1977.

[17] R. Zafarani and H. Liu. Social computing data repository at ASU. http://socialcomputing.asu.edu/, 2009.

[18] Y. Zhou and J. Davis. Community discovery and analysis in blogspace. In *Proceedings of the 15th international conference on World Wide Web*, page 1018. ACM, 2006.