Using Lists to Measure Homophily on Twitter

Jeon Hyung Kang
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
jeonhyuk@usc.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
lerman@isi.edu

Abstract

Homophily is the tendency of individuals in a social system to link to others who are similar to them and understanding homophily can help us build better user models for personalization and recommender systems. Many studies have verified homophily along demographic dimensions, such as age, location, occupation, etc., not only in real-world social networks but also online. However, there is limited research showing that homophily also exists when similarity is judged by topics of expertise or interests. We demonstrate the existence of topical homophily on Twitter using a novel source of evidence provided by Twitter lists. In this paper, we use LDA to extract topics from Twitter lists (a collection of user accounts created by some user that others can follow) and measure similarity between listed users based on the learned topics. We show that topically similar users are more likely to be linked via a follow relationship than less similar users.

Homophily is a strong organizing principle of social systems and has been used to explain human and social behavior. Homophily refers to the tendency of individuals in a social system to link with others who are similar to them rather than those who are less similar. The community structure homophily imposes on the social network may, in turn, through the processes of influence (Christakis and Fowler 2007) and selection (Crandall et al. 2008) cause linked individuals to become even more similar. Over time, preferential linking will structure the network in such a way as to make the behavior of individuals (Lerman et al. 2011) and even future friendships (Liben-Nowell and Kleinberg 2007) more predictable.

Understanding homophily can help us build better models for user/item recommendation systems and web personalization services by taking into account users' similarities and their social behaviors. Existence of demographic homophily, that is homophily based on demographic characteristics, is well established (Feld 1981; Kossinets and Watts 2009). (McPherson, Smith-Lovin, and Cook 2001), for example, cites over a hundred studies that support homophily along multiple dimensions, such as race, ethnicity, sex, gender, age, religion, education, occupation, abilities, beliefs, aspirations, and so on. Less empirical evidence exists for homophily in online social networks in which individual interactions are not constrained by geographic and organizational proximity and are instead based on shared interests or expertise. In an online social network of Twitter, for example, one is more likely to find a Semantic Web researcher who is linked to another Semantic Web researcher than to an app developer and vice versa, despite similar demographic characteristics of the two groups of users. One challenge to demonstrate homophily is to define a metric that properly accounts for topical similarity. (Singla and Richardson 2008) used the categories of search queries issued by users, in addition to their demographic characteristics, to measure similarity and demonstrated that people who talk to each other via instant messaging are more likely to be similar than a random pair of users. Others (Weng et al. 2010; Schifanella et al. 2010) found that linked social media users share topical interests and tagging vocabulary, and (Wu et al. 2011) found homophily within categories, with celebrities tending to follow other celebrities, bloggers other bloggers, and so on.

In this paper we show homophily on Twitter using *Twitter lists* as a novel source of evidence for topical similarity. In addition to broadcasting short messages, called tweets, registered Twitter users can follow accounts of other users to receive their tweets. Twitter introduced lists to help users manage the friends they follow. A *list* is created by some user, referred to as the *curator*, who names it and adds up to 500 members to it. A curator can create up to 20 lists. Other users can then *subscribe* to the list to see tweets from list members without having to follow them directly. Essentially, Twitter users categorize others by tagging them with list names. By applying topic modeling techniques to lists, we find the reduced dimension topic space which serves as a basis for measuring similarity between list members. We find that topically similar users are more likely to be linked via a follower relationship than less similar users.

In Section "Twitter Lists" we describe our data collection methodology and properties of Twitter lists. Just like tags that are used to annotate resources in social bookmarking sites, list names are used to categorize user accounts on Twitter along multiple dimensions and have a long tailed frequency distribution. Unlike tags, however, lists add a new relational layer to Twitter data, since they are used to indirectly follow users. In Section "Topical Homophily on Twitter",

we use Latent Dirichlet Allocation to learn the topic distribution of Twitter list members in our sample. We define a similarity measure based on these topics and empirically demonstrate existence of homophily. Our work demonstrates the potential of Twitter lists for numerous applications, including discovering communities of shared interests, experts on particular topics, categorizing people within subject matter directories, ranking popular or influential users, recommending interesting users or tweets based on a topic, and so on.

Twitter Lists

Twitter offers an Application Programming Interface (API) for data collection. Researchers have used various data collection strategies to overcome the biases of limited sample size exposed by Twitter through the API (Krishnamurthy, Gill, and Arlitt 2008; Wu et al. 2011). We collected a snowball sample of users and *lists* as follows. Starting with two initial *seed* users, we collected all lists, schematically shown in Fig. 1(a), that they subscribed to or were members of: a total of 260 lists. Next, we expanded the user layer, as shown in Fig. 1(b), based on the current lists by collecting all other users who are members of or subscribers to these lists. This yielded an additional 2573 users. In the next iteration, we expanded the list layers by collecting all lists that these users subscribe to or are members of. This raised the number of lists from 260 to almost 298K. In the last step, we collected users associated with these 298K lists, yielding 905K users. In addition, we also collected information about who these users were following. In total, the snowball sample contained 298K lists, and 2.3M users with 111M friendship, 10M membership, and 1.5M subscription links.

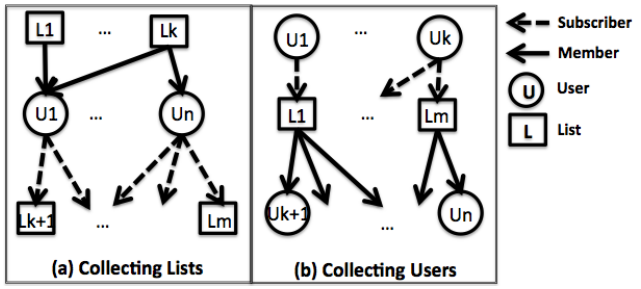


Figure 1: Schematic representation of snowball sample data collection. (a) We collected all lists that n users are members of (from L_1 to L_k) and subscribers to (from L_{k+1} to L_m). (b) Starting from m newly discovered lists, we collected all users (from U_1 to U_k) that subscribed to them or were members of (from U_{k+1} to U_n). We repeated steps (a) and (b) to collect another layer of lists and users.

As a specific example, one of the seed users `@jahendler` (Prof. Jim Hendler of RPI), a self-described "SemWeb guru, Web Science evangelist, general web geek," has been listed 107 times. The lists to which he has been assigned have names containing terms 'semantic' (31 lists), 'web' (30 lists), 'semweb' (22 lists), 'semanticweb' (18 lists), 'tech' (8 lists), 'technology' (6 lists), 'science' (6 lists), 'rpi' (5 lists), 'opendata' (4 lists), 'research' (4 lists), 'analytics' (3 lists), 'people' (3 lists), 'media' (3 lists), etc. These terms, and identities of other members of these lists, offer additional insights into `@jahendler`'s interests and expertise.

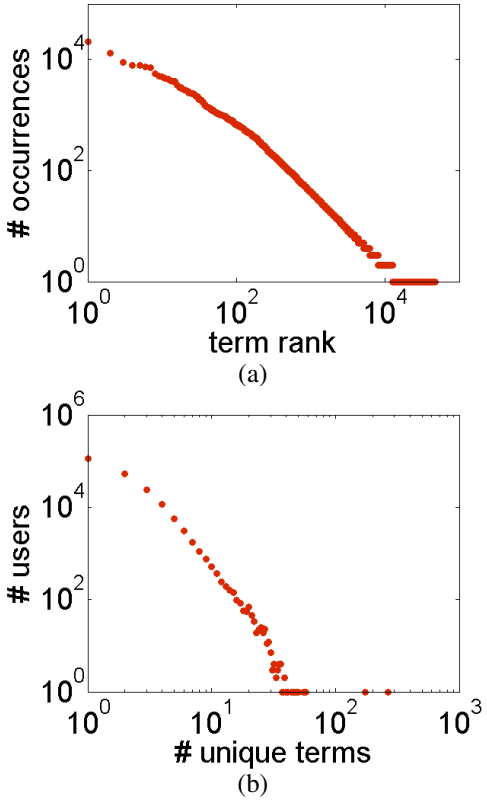


Figure 2: Term statistics. (a) Term frequency over the entire lists on log-log scale: terms are sorted by rank with the most frequent term ranked one. (b) Vocabulary size distribution of individual list curators.

Structural Analysis of Twitter Lists

The distribution of the number of friends and followers in the snowball data sample has a long tailed form, consistent with previous measurements (Java et al. 2007; Kwak et al. 2010; Weng et al. 2010; Bakshy et al. 2011). List subscription also shows long-tailed distribution with the numbers of list subscribers ranging from one to 70K.

In naming lists, Twitter users act very much like users of social tagging systems. In tagging systems, users' tag vocabulary (number of distinct terms they use) is broadly distributed and grows over time as users discover new interests and describe resources according to them (Golder and Huberman 2006). Figure 2 shows the frequency-rank distribution of terms in list names, as well as the vocabulary size distribution of list curators.[1] Both show a long-tailed dis-

[1]We decompose compound names into individual terms by to-

Term	Frequency	List name examples
new	21040	tech-news, news-magazined, world-news
tech	12940	all-about-tech, abbs-and-tech, digital-tech
twibe	8784	twibes-socialmedia, twibes-marketing
politics	7955	politics, political, news-politics, us-politics
media	7810	social-media, media, news-media
celeb	7266	celebs, celebs-i-follow, faves-celebs
people	7025	people, famous-people, funny-people
design	5522	design, web-design, designers
celebrity	5046	celebrities, celebrity-tweets
social	4853	social-media, social-networking
list	4599	my-favstar-fm-list, my-list
funny	4210	funny-people, funny-stuff, funny-folks
web	4050	web-development, web-2-0, web-tech
technology	4016	science-technology, technology-news
science	3468	science-tech, science-space
my	3160	my-govluv-reps, my-twitlets
business	3102	business-leaders,business-marketing
famous	2955	famous-people, famous-folk, famous-ppl
follow	2825	fav-follows, follow-friday, my-followers
entertainment	2764	entertainers, entertain-me

Table 1: Top 20 most frequent terms in list names

tribution, with a few terms, twenty of which are shown in Table 1, used many times, while the vast majority of terms are used only infrequently.

Linguistic Analysis of Twitter Lists

In social tagging systems, tags are freely chosen by the user from an uncontrolled vocabulary and describe different aspects of the tagged resource (Golder and Huberman 2006). Similarly, terms in list names are drawn by users from their personal uncontrolled vocabularies to describe a variety of characteristics of listed users (members) including:

- identify the member: e.g., art, business, music, sport
- identify the type of member: e.g., guru, people, talk, quote, video, audio, pic
- identify its characteristics: e.g., inspiring, interest, develop, creative, innovative, good, influence, funny, stupid, famous
- specify some social behavior: e.g., friday follow, my-followers, follow-back, recently-followed-me
- provide refining categories: e.g., list-1, list-2, list-3, wednesday follow, thursday follow, friday follow

To analyze the terms in Twitter lists, we first normalized terms by tokenizing list names (on hyphens) and stemming individual terms. Our data set contains 298K lists with the total 462K terms, of which 48K are unique. The most frequent term was "news", which occurred 21K times, suggesting the important role mass media plays even on Twitter (Wu et al. 2011). In addition, in our data set people curate or subscribe to lists not just for keeping track of funny things (i.e., terms such as "funny", "interesting") but also to be inspired (23rd ranked "inspire" and 29th ranked "creative"). We manually categorized most frequent Twitter list names in Table. 2. Even though we started from two computer science researchers, our final snowball data sample contains users who subscribe to or are members of a wide variety of lists.

kenizing list names on the hyphen.

Category	Sub-category	List Name Examples
Common interest	Information	travel-deals, travel-agents, shop-and-savings, pics from Space, food-truck, nyc-food, health-wellness, wine-lovers, wine-spirits-cocktails
	Fan	we-love-justin-bieber, i-love-gaga, jackson-fan, justine bieber, celebs, celebrities, favstar, Sports-fan, lakers-fan
	Media	Social-media, it-media, film-media, wine-media-marketing, LA-media, blogs, tech-blogs, design-blogs
	Hobby	books, music-related, music-artistis, sports-general, art-design-photo, design-photo, my-favstar-fm-list, movies, movie-people, tv-shows
	News	news-politics, news-and-politics, business-news, tech-news, marketing, internet-marketing, online-marketing, it-news, it-tech, info
	Leaders	politicians, marketing-gurus, business-leaders, tech-people, music-artists, art-artists, science-writers, food-bloggers, travel-bloggers, ceo-founders, professors, designers
	Fun	most-liked, stuff-i-like, things-i-like, humor-comedy, celebrity-gossip, laughs-and-gossip, sports-entertainment, video-games, movies
	Art	web-design, web-development, web-tech, design, graphic-design, design-resources, web-designmers, art-group, art-design
Business relation		business-contacts, work-contacts, marketing-contacts, job-search, job-postings, jobs, employers, employers-jp, employment-tweets
Common personal traits		los-Angeles, tokyo, san diego, people-like-me, most-liked, people-i-like, things-i-like
Social Purpose		local-friend, who-my-friends-talk-to, mutual-friend, twitter-friends, conversation-list, conversations-and-chats, family, friday follow
Self-Motivation		inspirational-quotes, inspire-motivate, creative-thinkers, creativity-innovation, innovators, innovators-influencers, great-quotes, influencers, tech-influencers

Table 2: Categorization of Twitter list names

Direct and Indirect Following

Twitter users broadcast messages to their followers, and in turn follow the updates of others. The following behavior on Twitter has been studied by many researchers (McPherson, Smith-Lovin, and Cook 2001; Wu et al. 2011; Kwak et al. 2010; Huberman, Romero, and Wu 2008; Weng et al. 2010; Krishnamurthy, Gill, and Arlitt 2008; Mislove et al. 2007). In addition to *directly* following others on Twitter, lists allow users to *indirectly* follow them, since by subscribing to a list, a user will receive updates from list members even if she is not directly following them (the user has to click on the list to get the updates).

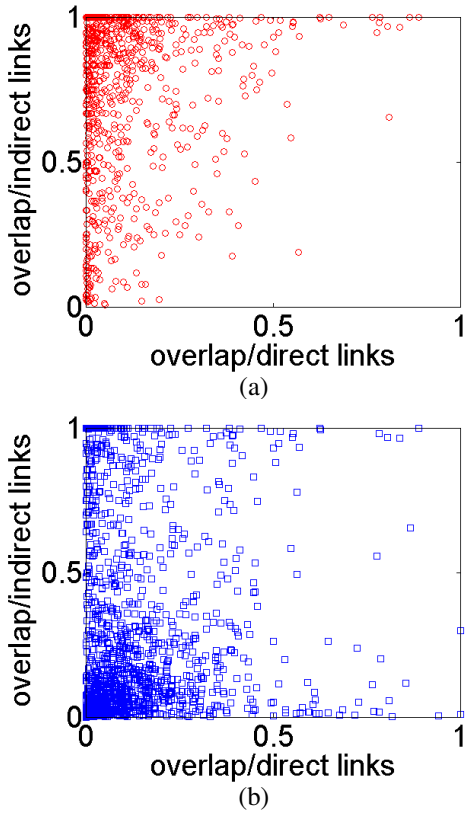


Figure 3: Direct (via follower links) vs indirect (via lists) following comparison for (a) list curators and (b) list subscribers.

How much overlap exists among those users are following directly and indirectly? We compute the overlap quantitatively. User can follow others through n direct (follower) links (DIR_i=u_1, u_2, ...,u_n) and m indirect (list) links ($INDIR_i = u_1, u_2, ..., u_m$). Let k be the number of common links ($OVERLAP_i$=u_1, u_2, ..., u_k). Figure 3 shows the degree of overlap relative to the numbers of people followed directly and indirectly, with each point having values (k/n, k/m). Figure 3(a) shows this distribution for list curators, and Fig. 3(b) for list subscribers. In both of cases, people tend to indirectly follow fewer than 50% of the people they follow directly. List curators tend to indirectly follow the users they are already following, while subscribers tend to use lists to indirectly follow new people whom they are not already following. This indicates that list curators tend to use lists as a means to categorize people they already follow, while subscribers use them as a source of new information.

List Curating Behaviors

Two different behaviors were observed in social tagging systems: categorizing and describing (Körner et al. 2010). Describers use a variety of tags to describe an object, while categorizers use one or at most a few tags to exactly place the object within some categorization scheme. Twitter lists are similar to social tagging systems in a sense that lists represent a non-hierarchical assignment of objects (in this case people) to categories. We observe the two behaviors also on Twitter. Some users (describers) generate lists with similar membership but different names, while others (categorizers) generate disjoint lists with different users. Tagging pragmatics can be measured by vocabulary size, tag/resource ratio, average tags per post, and orphan ratio (Körner et al. 2010). To analyze different list curating behaviors, we compute $DC_i = \sum_{j=1:L} (j \times n_{ij})/(N_i \times L_i)$, where L_i is the number of lists that user i curates, N_i is the number of unique members in the lists that user i curates, and n_{ij} is the number of users appeared in j different lists that user i curates. If user i creates N lists with the same L members, DC_i will be 1 and we will conclude that i is a describer. On the other hand, if user i creates N disjoint lists, DC_i will be close to 0, and we will conclude that the user is categorizer. This gives more credit to categorizers who have more list if both users have totally disjoint lists.

We selected 2,243 users who curate between 6 to 20 lists (note that 20 is the maximum number of lists one can curate) and computed their DC scores, which are shown in Fig. 4(a). The figure illustrates that people tends to create multiple lists using similar set of users rather than create disjoint set of lists for exclusive categorization. List curators tend to follow many people. Figure 4(b) shows the degree comparison between Twitter list curators and non-curators. Blue circles represent curators, and red rectangles represent user who have not curated any lists. Curators tend to follow more users than they themselves are followed by, and many of them have no followers. Compared to most of curators, non-curators have more followers.

Learning the Topic Model of Twitter Lists

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a popular method for automatically extracting a compressed description of a document corpus. LDA is a completely unsupervised model that views documents as mixtures of topics represented as a K dimensional random variable θ. Each topic is represented as a probability distribution over words. Given a collection of documents, it is possible to learn the latent topics that best explain the words observed in the documents. In this generative model, a document is generated by first picking a topic distribution θ from the Dirichlet prior, and then using the document's topic distribution θ to sample latent topic variables z_i. LDA makes

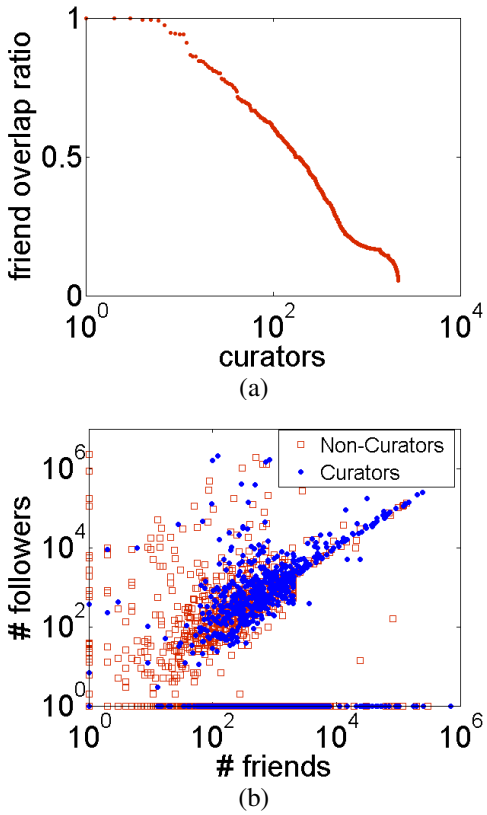


Figure 4: (a) DC scores of 2,243 users who curated the most lists. (b) Degree comparison between list curators (blue circles) and non-curators (red squares). X-axis shows the number of friends a user follows, and y-axis the number of followers she has.

the assumption that each word is generated from one topic, where z_i is a latent variable indicating the hidden topic assignment for word w_i. Probability of choosing a word w_i under topic z_i, $p(w_i|z_i;\beta)$, is different for all documents.

We use LDA to learn the hidden topics of Twitter lists. We view each list member as a "document," which is represented as a mixture over 200 topics, and topics as distributions over terms in list names that they are members of. The corpus consists of 140,231 users who have been listed at least ten times.

If two users are assigned to Twitter lists on similar topics, their topic distributions should be similar. Figure 5 shows the topic distributions of nine popular Twitter users, with six highest probability (stemmed) terms for selected topics listed under the figure. Topic distributions of `@BarackObama` and `@whitehouse` are both peaked at topic 61 (politics, politician, government, etc.). Accounts of mass media news sources `@cnnbrk` and `@nytimes` are both peaked at topic 50 (news, information, media). Note that `@BarackObama` and `@whitehouse` also have a substantial probability mass at this topic. In the same context, `@techcrunch`, `@google` and `@mashable` have a peak at topic 52 (technology, web, science, etc.). Topic distributions of `@DalaiLama` and `@BillGates` are peaked at

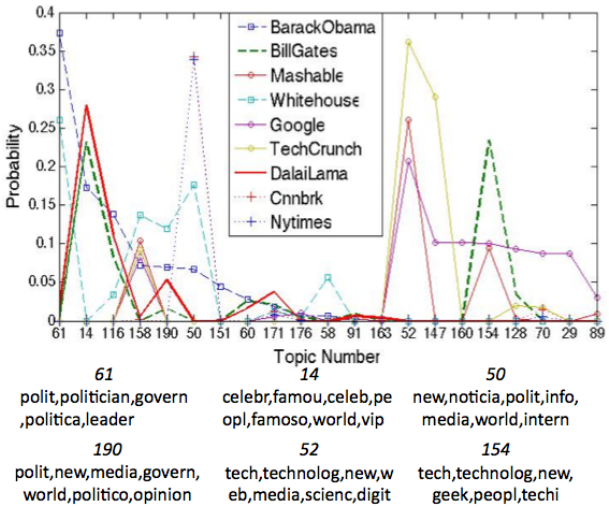


Figure 5: Topic distribution θ of nine members (for simplicity, distributions over 20 topics only are shown), and most probable words in six of the topics.

topic 14 (celebrity and famous people). However, their other topics are different in that `@DalaiLama` has the second largest peak at topic 190 (politics and government) while `@BillGates` has the second largest peak at topic 154 (technology, news and geek). Qualitatively at least, Twitter lists seem to capture topical similarity between list members.

Topical Homophily on Twitter

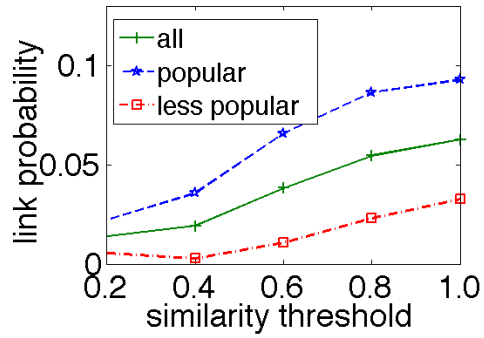
Does topical homophily exist on Twitter? In other words, are users who are more topically similar to each other, such as `@DalaiLama` and `@BillGates` or `@BillGates` and `@google`, more likely to be linked in the follower graph than users who are dissimilar, e.g., `@DalaiLama` and `@google`? We study this question using Twitter lists as evidence to measure topical similarity. Specifically, we calculate the similarity between two users who are list members using Jensen-Shannon divergence (Lin 1991) of their learned topic distributions.

Empirical Results

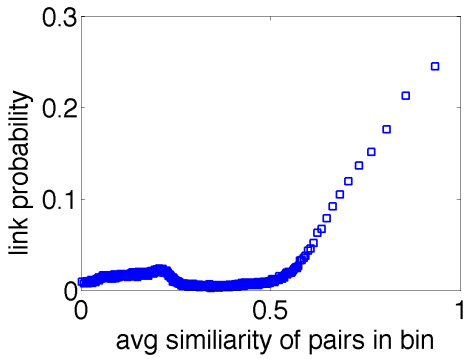
We performed two experiments to answer the question "are more similar users more likely to be linked ?" We computed user-to-user pair similarities by subtracting Jensen-Shannon divergence from 1, so that similarity score ranges from 0 (most dissimilar) to 1 (most similar). We say that users are linked if either a friend or a follower relationship exists between them. In other words, a link exists between users a and b if either user a follows user b or user b follows user a.

In the first experiment, we analyze the relationship between the likelihood of a link between pairs of users and their topical similarities. For each user in our data set who is a member of at least ten lists, we computed pair-wise similarities with the remaining 140K list members and binned together the top 2,000 pairs whose similarity is below some threshold. Note that each pair is either linked or not with certain similarity value, and we compute the likelihood of a link

between pairs by binned together 2,000 pairs. We used five different threshold values (1.0, 0.8, 0.6, 0.4, and 0.2) to bin pairs of list members. Next, we computed the percentage of linked pairs in each bin. Figure 6(a) shows that probability of a link increases steadily with similarity threshold; therefore, topically similar users are more likely to be linked. To make our findings concrete, we presented continuous trends on a more granular level. We paired 4,980 members, sorted them by similarity and divided evenly between 1,000 buckets. Figure 6(b) shows that 24.5% of pairs are linked when their average similarity is high (similarity $< 93\%$), while only 2.3% of pairs are linked when their average similarity is low (similarity $< 1\%$). As similarity of pairs increases, the probability of a link also increases. Note that only 1.098% of the 2.4M pairs are linked by either friends or followers relationships. Also, 4,980 members are connected to, on average, 32K followers (from 699 to 7.4M) and 6,208 friends (from 1 to 699K) in the whole Twitter follower graph, and 306 followers average (from 0 to 26K) and 1,151 friends average (from 0 to 30K) from 140K list members.



(a)



(b)

Figure 6: Probability of a link between list members. (a) Probability of a link among top 2,000 pairs whose similarity is below some threshold. Users are divided into two classes: popular and less-popular members. (b) 2.4M pairs of list members sorted by similarity score in decreasing order and divided evenly between 1,000 buckets. Each symbol represents probability of a tie of each bucket with average similarity value in x-axis.

These results show that similar users are more likely to be linked than dissimilar users. However, is this effect produced by homophily, or some other phenomenon, such as assortativity on Twitter? Assortativity measures the preference of popular (many followers) users to be linked to other popular users. To verify that the observed trends are not caused by assortativity, we divided list members into two categories of equal size: popular and less popular members. Popular members have on average 306K followers, while less popular members have on average 1K followers. We repeated the analysis above, dividing pairs of users into bins based on similarity threshold and measuring average probability of linking within each bin. Results, shown in Figure 6, show that the probability of a link increases from 2.22% to 9.26% for popular users, while for less popular users, probability of a link increases from 0.54% to 3.26%. Even though linking probability is different in the two classes of users, possibly due to varying visibility or accessibility of their accounts, in both classes the probability of a link increases monotonically with similarity. We believe that homophily is the remaining explanation for this trend.

Related Work

Homophily is an well-researched topic in social science. Homophily, which describes the propensity of similar individuals to link at a higher rate, is an important factor in the evolution of social networks and the diffusion of ideas and behaviors on these networks. Many studies have verified that people associate with each other and communicate at higher rate if they are similar along demographic characteristics, such as race, ethnicity, sex, gender, age, religion, etc. (McPherson, Smith-Lovin, and Cook 2001; Leskovec and Horvitz 2008; Kossinets and Watts 2009). Demographic homophily in social media sites such as Twitter has been demonstrated by several researchers, e.g., (Mackinnon 2006; Kwak et al. 2010) found homophily on Twitter based on users' age and country of residence. However, there is only limited research demonstrating that homophily also exists when similarity is judged by users' expertise or topics of interest. (Singla and Richardson 2008) showed that people chat with each other more often when they share interests. (Weng et al. 2010) found that users who reciprocate friendship links on Twitter tend to share topical interests. (De Choudhury et al. 2010) investigated the interplay between homophily along diverse user attributes and the information diffusion process on social media.

(Crandall et al. 2008) studied the interplay between similarity and network ties among Wikipedia editors and found that rising similarity predicts future interactions. (Gilbert and Karahalios 2009) presented a predictive model that maps social media data to social tie strength using thirty two variables including demographic, emotional, structural (e.g., number of mutual friends, friends of friends), and other features. We demonstrate the existence of topical homophily on Twitter using a novel source of evidence provided by Twitter lists. Unlike other studies that rely on users' demographic features, or features of content they create, to compute similarity, we use labels created by other users to categorize the users in question. These labels serve as a basis for calculating topical similairity. We show that users who are more

similar are more likely to link to each other via a friend or a follower relationship than users who are less similar. Unlike other studies, we also studied the relationship between topical and structural similarity.

Conclusion

In this paper, we use Twitter lists to demonstrate homophily on Twitter. Twitter lists are created by Twitter users to organize and categorize other users, and to indirectly follow topical accounts. As artifacts of human activity, Twitter lists offer a novel and rich data source for social data mining. We characterize statistical and linguistic properties of Twitter lists and show how they can be used to measure topical similarity between pairs of users. We demonstrated that Twitter users who are topically more similar are also more likely to be linked via a follower relationship than users who are less similar, and that this effect cannot be explained by other factors. Twitter lists allow us to explore two distinct properties of social networks: semantics and structure. In future work we will study how topical similarity affects the behavior of social networks, such as information diffusion.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In King, I.; Nejdl, W.; and Li, H., eds., *WSDM*, 65–74.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *J. Machine Learning Research* 3:993–1022.

Christakis, N. A., and Fowler, J. H. 2007. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine* 357(4):370–379.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, KDD '08, 160–168.

De Choudhury, M.; Sundaram, H.; John, A.; Seligmann, D. D.; and Kelliher, A. 2010. "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media?

Feld, S. L. 1981. The focused organization of social ties. *The American Journal of Sociology* 86(5):1015–1035.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proc. 27th Int. Conf. on Human factors in computing systems*, CHI '09, 211–220.

Golder, S., and Huberman, B. A. 2006. The structure of collaborative tagging systems. *Journal of Information Science* 32(2).

Huberman, B. A.; Romero, D. M.; and Wu, F. 2008. Social networks that matter: Twitter under the microscope. cite arxiv:0812.1045.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*.

Körner, C.; Benz, D.; Hotho, A.; Strohmaier, M.; and Gerd, S. 2010. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proc. 19th Int. Conf. on World wide web*, WWW '10, 521–530.

Kossinets, G., and Watts, D. J. 2009. Origins of Homophily in an Evolving Social Network. *The American Journal of Sociology* 115(2).

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *WOSP '08: Proc. first workshop on Online social networks*, 19–24. New York, NY, USA: ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *WWW '10: Proc. 19th Int. Conf. on World wide web*, 591–600. New York, NY, USA: ACM.

Lerman, K.; Intagorn, S.; Kang, J.-H.; and Ghosh, R. 2011. Using proximity to predict activity in social networks.

Leskovec, J., and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, 915–924. New York, NY, USA: ACM.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci.* 58(7):1019–1031.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory* 37:145–151.

Mackinnon, I. 2006. Age and geographic inferences of the livejournal social network. In *In Statistical Network Analysis Workshop*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proc. 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, 29–42.

Schifanella, R.; Barrat, A.; Cattuto, C.; Markines, B.; and Menczer, F. 2010. Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, 271–280. New York, NY, USA: ACM.

Singla, P., and Richardson, M. 2008. Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In Davison, B. D.; Suel, T.; Craswell, N.; and Liu, B., eds., *WSDM*, 261–270. ACM.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In *Proc. 20th Int. Conf. on World wide web*, WWW '11, 705–714.