2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Unveiling Hidden Patterns to Find Social Relevance

Enkh-Amgalan Baatarjav and Ram Dantu
Department of Computer Science and Engineering
University of North Texas
Denton, Texas, 76203, USA
Email: {eb0050, rdantu}@unt.edu

***Abstract*—Twitter is both a useful social networking device and an incredible marketing tool. However, it is also a venue for dangerous stalkers and a sub-world of internet users that most people would intuitively avoid if seeing them in real life. It would improve Twitter's safety to have filters available which would allow users to select an audience for their status updates[1] without being forced into changing their profiles to a private setting. The hypothetical filter, or model, studied in this paper was based on two particular attributes: activity correlations and vocabulary similarities between users and followers. If implemented, this model would restrict the availability of status updates to an automatically generated group of socially relevant followers. The result of this study shows that both of the attributes can be used to define social relevance; however, it was found that activity patterns have better predictive capabilities than correlating vocabulary usage between users and followers.**

***Index Terms*—Social Relevance, Online Social Networks, Twitter, Privacy.**

I. Introduction

Since the launch of the first online social networking site (OSN) in 1997, SixDegrees, there have been a wide range of OSN sites targeting different interests[1]. We can find OSN sites for making connections in business, dating, photo sharing, video sharing, music broadcasting, microblogging, mainstream networking, bookmarking, etc. These are just some of many OSN sites that build their themes based on their users' social network. Today, two of the most popular OSN sites are Facebook and Twitter, with Facebook having more than 500 million active users and Twitter having more than 105 million registered users[2].

OSN sites have revolutionized the way we communicate and share information. For instance, 5,000 status updates were posted on Twitter in 2007. However, the average number of status updates per day increased dramatically during the next few years. The 300,000 status updates in 2008 rose to 2.5 million the following year, and after that 35 million at the end of 2009. But by just the beginning of 2010, there were 50 million status updates. In other words, there were 600 status updates being posted every second. This number can only increase as more and more people adopt this communication medium. 96% of the Millennial generation, an amount estimated to be 80 million people, have already joined a social networking site[3]. A study conducted between October 2008 and February 2009 by Inside Facebook[4] shows that the fastest growing demographic on Facebook is women who are 55 years of age, with the growth rate being 175.3%. This is an indication that OSN sites are being embraced by many different age groups as a way to communicate.

Since this influential and fast growing new way of commination is not fully understood, it is continually being researched to unveil its potential.

[1]On the Twitter network, status updates are called Tweets.

II. Application in Real Life

Existing online social networking (OSN) sites are mostly used as an information broadcasting service. Users write a short sentence answering the question, "What's on your mind?"[5] or "What's happening?"[6] (This short sentence will be referred to as a status update from now on). A user's status update is then broadcasted to their social graph and potentially to the public as well. Depending on the users' privacy settings and the policies of social network sites, status updates are distributed differently.

On Facebook, users can broadcast their status updates to five different sets of users: everyone, friends and networks, friends of friends, only friends, or custom (a specific set of people selected at the time of the update). Currently, the default privacy setting is on everyone, which means that anyone on the Internet can see the status update if one searches for the user's profile. In addition, public status updates are searchable outside of Facebook[7]. Many people may not realize that their default privacy setting allows for public viewing on the Internet.

On the other hand, Twitter uses a binary approach for setting the privacy of status updates. Status updates can be either public (everyone on the Internet can see them) or private (only people who are are given permission to follow can see). However, like Facebook, the default privacy setting is on public. Both public updates on Facebook and Twitter are available through Google Real-Time search, which was launched on Dec 7, 2009[8].

Privacy may be an issue for the person posting a status update, but looking at status updates from the receivers' perspective presents a different set of problems. The sheer volume of status updates posted in a single day can be quite overwhelming. On Facebook, these status updates come as part of the Live Feed. In addition to status updates, it contains updates of neighbors' activities, such as the usage Facebook applications, changes on profiles and relationship statuses, etc. Average Facebook users have 130 friends[9] and create 90 pieces of content each month, which means that the average

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

user receives about 390 status updates on his/her news feed daily.

Many of these feeds are not relevant enough for a user to want to spend a lot of time reading. Additionally, relevant updates from closer friends are easily overlooked since there is such a large volume of information to sort through due to the "last in last out" model of displaying updates. Therefore, having a privacy management layer that helps to broadcast posts to a specific, relevant audience, instead of mass broadcasting, has a ripple effect on the recipients' quality of news feeds or received posts.

III. Problem Definition

Even though online social networking sites have been thoroughly integrated into our everyday lives, some people are still skeptical about sharing their personal information online because of insufficient privacy and security features. The main goal of our research is to create a privacy management system that only allows for a specified set of a user's friends/followers to be able to receive user's status updates.

One problem with Twitter is an assumption that all "friends" within a social network are equally close or relevant. This assumption precludes the fact that users have different levels of interaction with these "friends".

Another downfall of using Twitter is that some users post sensitive content in their status updates, as they can contain information about the users' daily and future activities. For example, a user's location is sensitive information because it can expose the habitual route of the user to the potentially hazardous and irrelevant public [10] [11].

A. Main Contribution

Existing social networking sites lack sufficient privacy management, as they are unable to define social relevance among friends/followers and broadcast status updates. Thus, in this paper, we explore just how to find a socially relevant set of a user's followers. However, the question that remains is which attributes help to define social relevance?

In order to create an effective model that blends seamlessly into the framework of real social networking sites, the attributes used in the model need to be available in many social network sites. An example of these are status updates and the timestamp of status updates, since most social networking sites offer these features. Two universal attributes considered in this particular study are activity correlation and vocabulary usage similarity.

IV. Background

Online social networking (OSN) sites have become an integral method of communication when connecting to family and friends. It also can facilitate the sharing and gathering of information with strangers, who may or may not have similar interests. Practically, however, this is not always the case. Thus, effectively managing privacy on social networking sites has started gaining more and more attention from individuals in many fields of study, particularly computer science, information science, and sociology. Because the number of publications related to this topic has grown considerably, it is now somewhat challenging to do a thorough literature survey. Although this survey has been narrowed down to publications from 2007 and 2010, it is not, by any means, exhaustive. The current research can be divided into two main domains: privacy vulnerability analysis and privacy protection models.

A. Privacy Vulnerability Analysis in OSN

Average OSN users can be easily manipulated if they believe that OSN sites are entirely secure, and therefore would be more willing to post personal information [12]. A study done on the usage of privacy settings shows that a majority of Facebook users do not change their default privacy settings even though they are able to limit the visibility of their profile information from strangers [13]. Additionally, having a large amount of personal information available in easily harvestable environments like Facebook, MySpace, and Orkut can lead to social phishing attacks [14]. Phishers can impersonate the friends of victims or use personal and social information in other harmful ways.

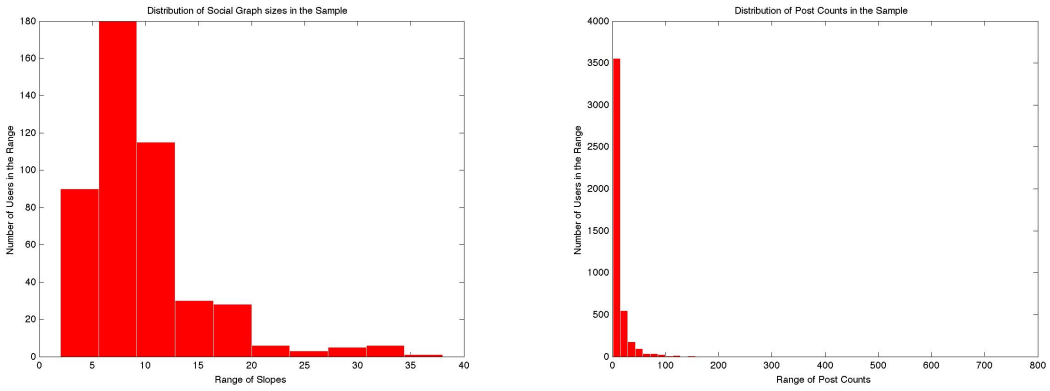
In a research study conducted by Krishnamurthy and Wills [15] on 3,851 randomly selected MySpace users, 1,600-1,700 Facebook users[2], and other OSN sites, the main privacy leakage sources came from default privacy settings, users' utilization of privacy-setting, and third-party applications.

Strater's and Richter's study [16] quantified college students' disclosure and privacy behavior on Facebook with the intent to develop a better future privacy system. Participants completed demographic surveys and a personal inventory (NEO-FFI). Each participant evaluated two other participants' profiles. The following is the highlight of the study: 77% of the participants had a publicly accessible profile. Even though the participants were aware of both the privacy concerns associated with Facebook and how to manually configure their own privacy settings, many participants still maintained the default setting, which is that the profile would be publicly accessible. The participants exhibited "all or none" approaches to disclosing their personal information.

Dwyer et al. [17] showed ways in which OSN site users control their online privacy and also how to generally improve the privacy management system. 222 members of the New Jersey Institute of Technology (NJIT) participated in this study, with 107 of them being Facebook users and 115 subjects being MySpace users. 18.9% of participants suffered a privacy incident, and less than half of them took counter measures to protect their privacy by changing their privacy settings. Their conclusion was that "privacy must be conceptualized as a quality of an online space, rather than as a collection of access settings to be managed by individual members."

Schrammel et al. studied [18] 850 online survey takers' information disclosure behavior and the correlation between information disclosure and demographic background. The study

[2]Users were sampled from 20 different U.S. regional networks and 20 different non-U.S. regional networks.



(a) Distribution of Social Graph Size

(b) Distribution of Post Count

Fig. 1. Distribution of post count and social graph size in the sample.

included different online communities: business networks (e.g. LinkedIn and Xing), social networks (e.g. Facebook and MySpace), content and media sharing networks (e.g. Flickr and YouTube), and social news and bookmarking sites (e.g. del.icio.us and Digg). The information disclosure behavior on different networks was analyzed based on nine common attributes among the different networks. Some of the highlights were that students are more willing to disclose their information on business, social, and social news and bookmarking networks, but not on content and media sharing networks, indicating that revealing information is directly related to the level of trust users have in the network.

Based on a dataset of 35,000 users, Lindamood et al. [19] argued that OSN users are vulnerable against attacks which use inference. They carried out an experiment based on 35,000 Facebook users, modifying the Naïve Bayes algorithm to classify a large amount of SN data. The algorithm took node traits and link structure as its inputs to predict a various user's political preference as either liberal or conservative. The experiment result was studied in four different cases: without removing any predictive trait and links, removing the 10 most predictive links and none of the traits, removing the 10 most predictive traits and none of the links, and removing the 10 most predictive traits and links. Their algorithm performed better than the traditional Naïve Bayes algorithm and Links only algorithm. It also performed better than the Details only algorithm in most cases.

Another inference study on 66,766 personal profiles on the Livejournal network was conducted by He et al. from the University of California in Los Angeles [20]. They used a Bayesian network approach to learn social relations on online SN, and they called it the Bayesian inference. The Bayesian inference outperformed the traditional Naïve inference. Their study of influence strength and society openness revealed both that personal attributes are highly correlated to the strength of the relationship between a user and the user's social graph, and that an effective way to protect privacy disclosure is to selectively hide friendship relations or friends' attributes.

B. Privacy Protection Model

Users' data, such as behavior, interest, and demographic, is very valuable for personalized systems and web applications. The data, when harvested, can be used to adapt to each user's personalized needs. Even though personalized systems are beneficial in providing relevant contents, targeted emails, and e-commerce, Internet users continue to express significant concerns about their privacy[21]. In the general context of the Internet, a lot of research and projects have been done to address the issue of protecting privacy. For example, the Platform for Privacy Preferences (P3P) project was created by W3C for privacy standard [22] by having proposed that user agents be integrated into browsers in order to check for the compatibility of users' privacy preferences and their website privacy policies. There have been several improvements made in the following research [23] [24] [25] [26].

A number of cryptographic approaches address the vulnerability of OSN users' privacy: Lucas and Borisov propose placing an encryption/decryption structure on social networks so that only users in the same social graph are able to decrypt information[27]. A variation of the cryptographic approach was also proposed in NOYB (None Of Your Business) by Guha et al.[28]. Anderson and colleagues[29] propose a client-server architecture using cryptography for social networking to ensure users' privacy.

In an extension of the previous work, *Inferring Private Information Using Social Network Data*[19] [30], Heatherly et al. proposed mitigating techniques to deflect inference-based attacks. The main argument for this study was that Facebook privacy configurations do not guarantee users' privacy, especially since a third party can predict undisclosed information of users based on just the friends of any user.

V. Dataset

The dataset used in this study was provided by Microsoft Research. It contains 1.3 million conversations gathered from the Twitter Public API from July 1, 2009 to August 27, 2009, including 477,045 unique conversations and 296,486

Twitter users. Conversations were built by following the ⟨in_reply_to_status_id⟩ tag from the API.

Due to limited resources, the stratified sampling approach was applied using ten different strata. The following steps were taken in order to sample a smaller portion of the data: first, the total number of users were grouped into a strata based on the number of conversation partners they have. Second, the strata was ordered in descending order, with the two 10% extremes from the top and the bottom being removed since an assumption was made that the two extremes of the strata were noise. Third, 40% of each strata was randomly sampled for the study, since it is thought that this amount can properly represent the data set. The distribution information of the sample is shown in Fig. (1). Fig. (1(a)) indicates the distribution size of users' social graph, and Fig. (1(b)) indicates the distribution of post count in a histogram with 50 bins.

VI. Methodology

Although some attributes that define social relevance are rather obscure, a careful analysis of just how social interactions are formed has revealed hidden patterns that can be integrated into a model. These attributes were discovered by analyzing the activity and vocabulary usage between users and their followers. First, followers were ranked based on social relevance, which is defined by the number of status updates sent from users to followers. This ranking is referred to as the base rank. Second, the activity patterns for users and their followers were calculated. For vocabulary usage, the same steps were applied by ranking followers based on the base rank and calculating the vocabulary similarity between users and their followers' vocabulary words. The relationship is then calculated by quantifying the activity correlation and their base ranks, and the vocabulary usage similarity and their base ranks.

A. Activity Pattern Analysis

Activity patterns are analyzed in two time domains: hour and week, and each one is divided into 24 and 7 sub time intervals, respectively. An example of this can be seen in Fig. (**??**). On each time interval, the average number of status updates is calculated to represent the activity pattern.

For the next step, we interpolate each activity pattern with degree of six and give a score to each follower based on how closely his/her activity pattern matches with his/her user's activity pattern. Finally, we rank the followers based on the score and compare the ranking with the social relevance ranking.

Having a positive correlation between activity pattern ranking and social relevance ranking indicates that we can use similarity score of users' and followers' activity patterns to define their social relevance.

Our architecture is shown in Fig. (3) consisting of three parts: analyzing activity pattern, activity correlation, and ranking based on the correlation. In the activity analysis, we analyze the activity level of Twitter users in two different time domains: hour and week. In the hour domain, we calculate

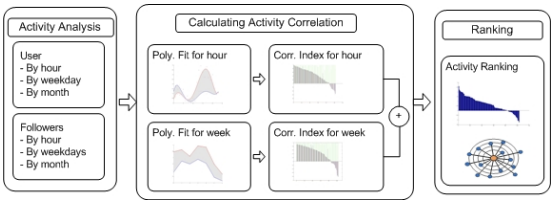


Fig. 3. Architecture is consists of three parts: activeness analysis, calculating activeness correlation, and ranking activeness correlation.

the average status updates for each hour of the day. The week domain contains the average status updates for each day of the week

In the next part, we calculate the correlation index for each follower in hour and week domains. To calculate the correlation index, we use an interpolation method, the polynomial model with a degree of six to fit the data points:

$$p(x) = \sum_{i=1}^{n} p_i x^{n+1-i}, \tag{1}$$

$n+1, p$ denotes the order ($n = 6$) and polynomial, respectively. We believe that the model fit well for our architecture because our data points are few and relatively simple. The polynomial model is relatively flexible enough to fit a limited number of points. Our experiment concludes a polynomial degree of six fits well for our data set. Using higher degrees introduces wild oscillation to the polynomial model and cause unfavorable result. After fitting the polynomial functions for both users' and their followers' data points, we can calculate the integral difference id of the functions, which describe the similarity of the functions based on the area under the curve:

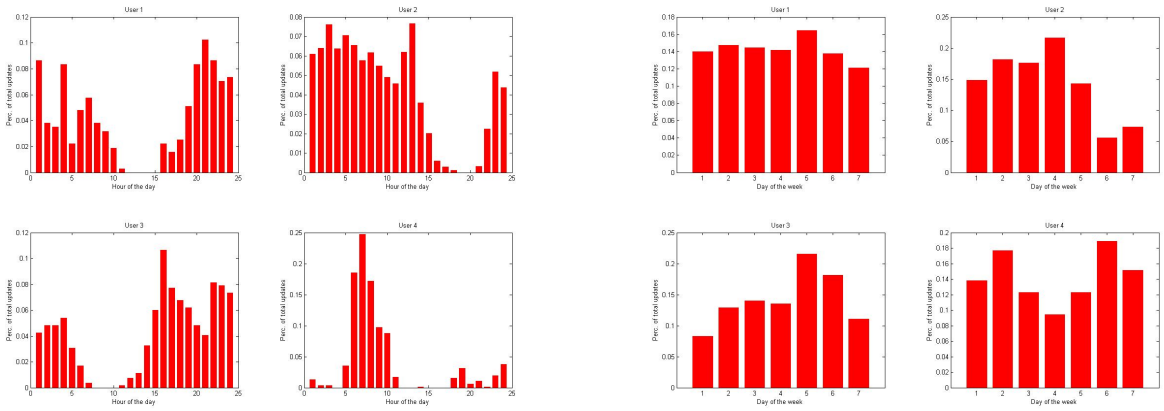
$$id_{u,f} = \int_1^N (p_u(x) - p_f(x))dx, \tag{2}$$

The number of intervals in each time domain are denoted by N. Hour and week domains have 24 (number of hour in the day) and 7 (number of days in the week) intervals, respectively. Polynomial function $p_u(x)$ represents a user's data points, and $p_f(x)$ denotes the polynomial function of follower's data points. In addition to finding id_{uf}, our architecture integrates the correlation coefficient to find a linear association between two polynomial functions $p(x)$. The correlation coefficient shows the strength and linear relationship of the user's and the follower's activity levels, which is denoted by the covariance $cov(u, f)$:

$$cc_{u,f} = \frac{cov(u,f)}{\sigma_u \sigma_f} = \frac{1}{N-1} \sum_{i=1}^{N} (\frac{u_i - \bar{u}}{\sigma_u})(\frac{f_i - \bar{f}}{\sigma_f}), \tag{3}$$

σ_u, σ_f denotes the standard deviation of the user's and the follower's activity level, respectively. The result of cc_{uf} will always be a value between -1 and 1. The result is interpreted as follows:

- If $cc_{u,f} = 0 \Rightarrow$ no correlation between u and f
- If $cc_{u,f} = 1 \Rightarrow$ strong correlation between between u and f



(a) The Hour Domain

(b) The Week Domain

Fig. 2. Examples of activity profiles of four users across each time domain.

- If $cc_{u,f} = -1 \Rightarrow$ inverse correlation between between u and f

Equation 4 summarizes our model to calculate the correlation index ci between the user and his/her followers.

$$ci_{u,f} = cc_{u,f}(1 - \frac{id_{u,f}}{max(id_{u,*})}), \tag{4}$$

In order to simplify the model, we normalize the id by dividing the maximum difference of the area under the curves $max(id_{u,*})$, and subtract it from 1. The result of the correlation index can have a value between -1 and 1. Positive $ci_{u,f}$ indicates a close activity pattern correlation between u and f. On the other hand, if the user and his/her followers have different activity levels, we have a negative result from $ci_{u,f}$. We apply our model Equation 4 to each time domain, and apply vector addition on the results to get the total correlation index $tci_{u,f}$ between the user and his/her followers:

$$tci_{u,f} = ci_{u,f}^{hour} + ci_{u,f}^{week}, \tag{5}$$

The final stage is to rank the followers based on their $tci_{u,f}$.

B. Vocabulary Usage Similarity

There are a number of approaches for measuring text similarities: the simple matching coefficient, the Jaccard Coefficient, the Tanimoto Coefficient, Correlation, and Euclidean distance. They all have their own benefits and disadvantages, depending upon which kind of dataset is considered, but all of the above methods can be applied to find text similarities. To make a valid argument of which one is the most efficient for the task at hand, every method should be compared after being implemented and run on the same dataset. In our study, we used the Euclidean distance approach after witnessing its decent performance with a smaller portion of the dataset. In this part, we analyze the content of the status updates. Our approach to finding social relevance from status updates is divided into three parts, as shown in Fig. (4). In Part 1, we apply some basic natural language processing techniques

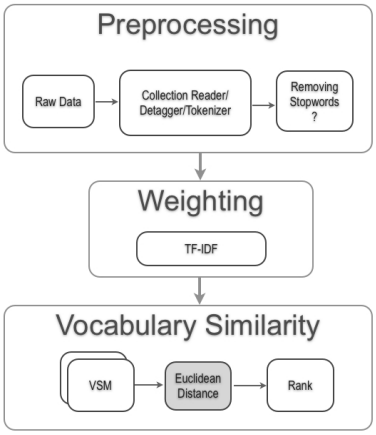


Fig. 4. Status updates similarity model is consists of three main parts: preprocessing, weighting vocabulary, calculating vocabulary similarity between users and their followers' vsm.

for cleaning the corpus and removing stop words from the data set, which reduce our computational time to complete by minimizing feature space. In addition, having stop words tend to reduce the performance of the system Fig. (4).

The next part of our analysis is to weight each vocabulary: not all words in the data set are evenly influential; Some word are frequently used in many status updates, and other words are used in only a set of users on their status updates. TF-IDF is a method used commonly to weight vocabulary. TF-IDF gives higher weight to vocabulary that is used frequently in a set of status updates and gives lower weight to vocabulary that is used frequently throughout the data set.

In the final part, we create vector space mode (VSM)

and calculate similarities of VSM between user's and his/her follower's. To create the VSM, first, we generate a bag of vocabulary for everyone by combine each user's and follower's status updates. Next, we generate dictionary of word, which is all vocabulary words in the data set. Finally, we calculate weighted average frequency of a word Eq. (6) in the dictionary for each user and follower using their bags of words, which is the vector space model (VSM) Eq. (7).

$$W_i = \frac{|F_i|}{|C|} \tag{6}$$

, where $|F_i|$ is frequency a word i and $|C|$ is total number of status updates.

$$V_{1...n} = w_1 * W_1, w_2 * W_2, ..., w_n * W_n \tag{7}$$

, where $|F_i|$ is frequency a word i and $|C|$ is total number of status updates. After finish calculating VSM for every one, we calculate vocabulary usage similarly between users and their followers by applying Euclidean distance function on the user's and the follower's VSM. If they use similar vocabulary similarity score will be high Eq. (SS).

$$SS(U_A, U_i) = Euclidean(U_A, U_i) \tag{8}$$

The final stage of this analysis is to rank the followers according to their similarity score. In the next section, we discuss our evaluation method and results of two approaches to find attributers that can define social relevance.

VII. Discussion of Results

Using Eq. (5) and Eq. (8), total activity correlation indexes and vocabulary similarity scores are calculated, and accumulative result of the sample data set is shown in Fig. (5). On each

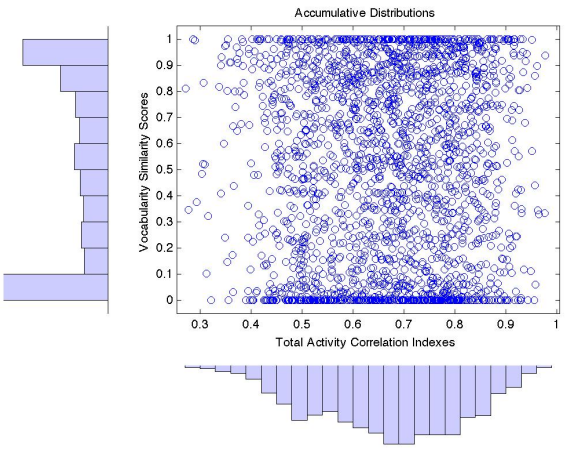


Fig. 5. Accumulative result of the total activity correlation indexes and vocabulary similarity scores. Histograms on *x-axis* and *y-axis* represent total activity correlation indexes and vocabulary similarity scores, respectively.

coordinate, value goes from zero to one: zero indicates the most dissimilarity between users and followers in respective attribute, and value of one indicates the most similarity. Even though distribution skews toward one on *y-axis*, there is large accumulation of points on zero. This indicates that a large number of followers who use similar vocabulary usage with their users, but about same number of followers have very different vocabulary usage from their users, as well.

Analyzing activity correlation indexes on Fig. (5) reveals that mean of the distribution is around 0.7: average followers have about 70% activity correlation with their users.

To evaluate the two attributes that were analyzed in this paper, the results of the rankings from the activity correlation and vocabulary usage similarity have to be compared against the true rank of social relevance. We define the social relevance by the number of messages that are exchanged between two different parties: the more messages that are exchanged, the closer the social relevance between the two parties is. We call this a base rank to which we are comparing our results.

The assumption of using number of interaction as social relevance or base rank is based on the studies conducted in area of psychology: socialemotional selective theory[32], selective optimization with compensation model[31]. Social interaction changes throughout lifespan from frequency of interaction with uniform distribution to positive skew distribution. The change was explained by social gain maximization, which is to increasing social and emotional gain and reducing emotional risk.

The following steps are taken in evaluation process, and Fig. (6) shows an example of a user with 38 followers. The figure shows calculated similarity score for total activity correlation index and vocabulary similarity score and fitted lines, as well. The variance of the total activity correlation index is much smaller than the variance of the vocabulary similarity score. Step 1, we order the followers based on their number of post

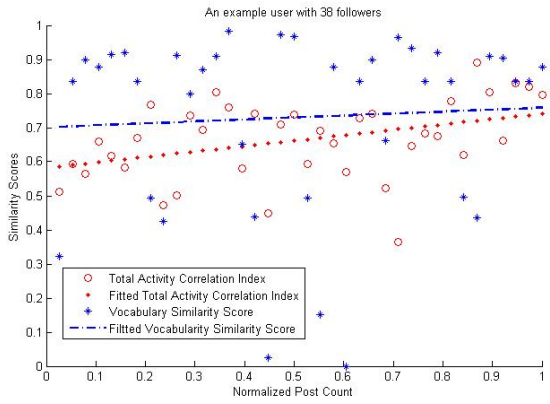


Fig. 6. Applying the linear regression model to fit a line on the similarity score reveals a tendency of the relationship between the calculated similarity score and the status update count.

messages they receive from their user on X-axis. Step 2, scores of activity correlation/vocabulary similarity are plotted. Step 3, we apply linear regression to model the relationship between base ranking and similarity scores or to show general trend

of the relationship: positive or negative. Positive relationship indicates if a follower has high similarity score, he/she tends to receive high number of status updates from user. Finally, we calculate the slope of the fitted line to observing the relationship trend.

Table (I) shows an overall result of both the activity correlation and the vocabulary similarity on the sample data set of size 4,518 and 2,93. Positive mean values in the table indicates that there is a positive correlation between the similarity score and the base rank. *The vocabulary similarity's low mean value and high STD on the table concludes that the activity correlation can give a better measurement than the vocabulary similarity for defining social relevance, even though it is intuitive to assume that people who use similar vocabulary have a high social relevance.*

TABLE I
OVER ALL RESULTS OF THE ACTIVITY CORRELATION AND THE VOCABULARY SIMILARITY

	Activity Correlation	Vocabulary Similarity
Mean	0.075042773	0.012846259
STD	0.180408439	0.450923205
Max	0.573787607	1.241298145
Min	-0.377833415	-1.108898947

The side by side comparison between activity correlation and vocabulary similarity is shown in Fig. (7). The distribution of the activity correlation slopes is narrower than the distribution of the vocabulary similarity slopes, which is good indication that activity correlation is a better approximation of social relevance than vocabulary similarity.

VIII. CONCLUSION

Online social networking has become popular among all generations, being used for both broadcasting and sharing information. Even though it is an efficient method of communication, there are still some privacy issues involving who on the social graph is able to view personal information. Existing social networking sites lack the privacy control needed to manage this beneficially. One approach is to find socially relevant followers with whom users want to share their information. In this preliminary study, we delve into non-obvious ways to define social relevance and we discover that the activity pattern and vocabulary similarities can be good representatives to define social relevance between users. The result of both attribute studies shows that there are positive correlations. In our future work, we will explore more ways to find social relevance.

ACKNOWLEDGMENTS

We would like to thank Mallory Schier, University of North Texas, for her technical editing expertise. This work is supported by the National Science Foundation under Grants CNS-0751205 and CNS-0821736.

REFERENCES

[1] D. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1-2, November 2007. [Online]. Available: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

[2] N. Bilton, "Chirp, twitters first developer conference, opens its doors," *The New York Times*, April 2010, [Online; accessed 26-April-2010].

[3] E. Qualman, *Socialnomics: How Social Media Transforms the Way We Live and Do Business*. Hoboken, New Jersey: John Wiley and Sons, Inc, 2009.

[4] J. Smith, "Fastest growing demographic on facebook: Women over 55," *Inside Facebook*, 2009.

[5] "Mainstream," http://www.facebook.com/, [Online; accessed 16-April-2010].

[6] "Microblogging," http://www.twitter.com/, [Online; accessed 16-April-2010].

[7] "Search Engine: Facebook Update," http://youropenbook.org/, [Online; accessed 20-June-2010].

[8] A. Singhal, "Relevance meets the real-time web," Google, Tech. Rep., 2009.

[9] "Facebook Statistics," http://www.facebook.com/press/info.php?statistics, 2010, [Online; accessed 16-April-2010].

[10] C. Haythornthwaite, "Strong, weak, and latent ties and the impact of new media," *The Information Society*, vol. 18, pp. 385–401, 2002. [Online]. Available: http://www.ingentaconnect.com/

[11] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociol Theory*, vol. 1, pp. 201–233, 1983.

[12] C. Dwyer, S. R. Hiltz, and K. Passerini, "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace," in *Proceedings of the Thirteenth Americas Conference on Information Systems*, August 2007. [Online]. Available: http://aisel.aisnet.org/amcis2007/339

[13] R. Gross, A. Acquisti, and H. J. Heinz, III, "Information revelation and privacy in online social networks," in *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. New York, NY, USA: ACM, 2005, pp. 71–80.

[14] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Commun. ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[15] B. Krishnamurthy and C. E. Wills, "Characterizing privacy in online social networks," in *WOSP '08: Proceedings of the first workshop on Online social networks*. New York, NY, USA: ACM, 2008, pp. 37–42.

[16] K. Strater and H. Richter, "Examining privacy and disclosure in a social networking community," in *SOUPS '07: Proceedings of the 3rd symposium on Usable privacy and security*. New York, NY, USA: ACM, 2007, pp. 157–158.

[17] C. A. Dwyer and S. R. Hiltz, "Designing Privacy into Online Communities," *SSRN eLibrary*, 2008.

[18] J. Schrammel, C. Köffel, and M. Tscheligi, "How much do you tell?: information disclosure behaviour indifferent types of online communities," in *C&T '09: Proceedings of the fourth international conference on Communities and technologies*. New York, NY, USA: ACM, 2009, pp. 275–284.

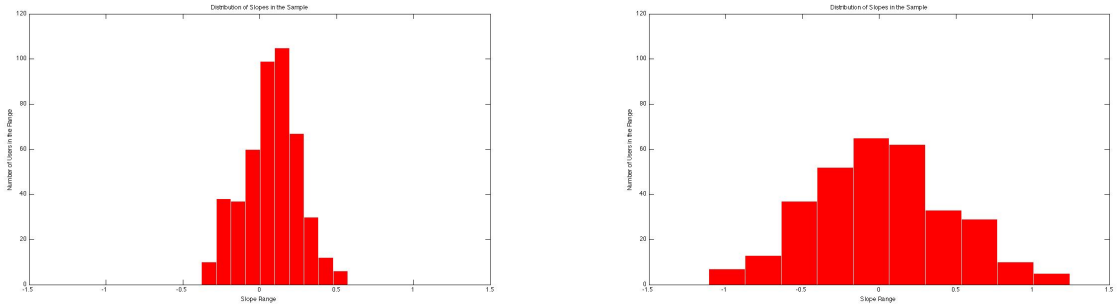
[19] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *18th International World Wide Web Conference*, April 2009, pp. 1145–1145. [Online]. Available: http://www2009.eprints.org/153/

[20] J. He, W. Chu, and Z. Liu, "Inferring privacy information from social networks," in *Intelligence and Security Informatics*, ser. Lecture Notes in Computer Science, S. Mehrotra, D. D. Zeng, H. Chen, B. Thuraisingham, and F.-Y. Wang, Eds. Berlin/Heidelberg: Springer-Verlag, 2006, vol. 3975, ch. 14, pp. 154–165. [Online]. Available: http://dx.doi.org/10.1007/11760146_14

[21] M. Teltzrow and A. Kobsa, "Impacts of user privacy preferences on personalized systems: a comparative study. in: Designing personalized user experiences for ecommerce." Kluwer Academic Publishers, 2004, pp. 315–332.

[22] L. Cranor, B. Dobbs, S. Egelman, and et al, "The platform for privacy preferences 1.1 (p3p1.1) specification," W3C, Tech. Rep., 2006. [Online]. Available: http://www.w3.org/TR/P3P11/

[23] S. Preibusch, B. Hoser, S. Gürses, and B. Berendt, "Ubiquitous social networks : Opportunities and challenges for privacy-aware user modelling," DIW Berlin, German Institute for Economic Research,



(a) Distribution of Activity Correlation Slopes

(b) Distribution of Vocabulary Similarity Slopes

Fig. 7. Side by side distribution comparison between activity correlation and vocabulary similarity.

Discussion Papers of DIW Berlin 698, 2007. [Online]. Available: http://ideas.repec.org/p/diw/diwwpp/dp698.html

[24] M. Teltzrow, S. Preibusch, and B. Berendt, "Simt - a privacy preserving web metrics tool," *E-Commerce Technology, IEEE International Conference on*, vol. 0, pp. 263–270, 2004.

[25] S. Preibusch, "Spontaneous privacy policy negotiations in pervasive environments," 2007, pp. 814–823. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-76890-6_7

[26] ——, "Privacy negotiations with p3p," in *W3C Workshop on Languages for Privacy Policy Negotiation and Semantics-Driven Enforcement*, 2006.

[27] M. M. Lucas and N. Borisov, "Flybynight: mitigating the privacy risks of social networking," in *WPES '08: Proceedings of the 7th ACM workshop on Privacy in the electronic society*. New York, NY, USA: ACM, 2008, pp. 1–8.

[28] S. Guha, K. Tang, and P. Francis, "Noyb: Privacy in online social networks," in *Proceedings of the First ACM SIGCOMM Workshop on Online Social Networks (WOSN)*, Seattle, WA, USA, August 2008. [Online]. Available: http://saikat.guha.cc/pub/cucs08-noyb.pdf

[29] J. Anderson, C. Diaz, Bonneau, and F. Stajano, "Privacy-enabling social networking over untrusted networks," in *Proceedings of the Second ACM SIGCOMM Workshop on Social Network Systems, 2009*. Barcelona, Spain: ACM, 2009.

[30] R. Heatherly, M. Kantarcioglu, B. Thuraisingham, and J. Lindamood, "Preventing private information inference attacks on social networks," UTDCS-03-09, Tech. Rep., 2009.

[31] C. Peterson, *Looking Forward Through the Lifespan: Developmental Psychology*. Pearson Education Australia, 2009. [Online]. Available: http://books.google.com/books?id=1wq2PwAACAAJ

[32] L. L. Carstensen, "Social and emotional patterns in adulthood: support for socioemotional selectivity theory." *Psychol Aging*, vol. 7, no. 3, pp. 331–8, 1992. [Online]. Available: http://www.biomedsearch.com/nih/Social-emotional-patterns-in-adulthood/1388852.html