Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

2-2010

Twitterrank: Finding topic-sensitive influential Twitterers

Jianshu WENG
Singapore Management University, jsweng@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Qi HE
Pennsylvania State University - Main Campus

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

Part of the Databases and Information Systems Commons, Numerical Analysis and Scientific Computing Commons, and the Social Media Commons

Citation

WENG, Jianshu; LIM, Ee Peng; JIANG, Jing; and HE, Qi. Twitterrank: Finding topic-sensitive influential Twitterers. (2010). *Proceedings of the Third ACM International Conference on Web Search & Data Mining: February 3-6, 2010, New York*. 261-270.

Available at: https://ink.library.smu.edu.sg/sis_research/504

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

TwitterRank: Finding Topic-sensitive Influential Twitterers

Jianshu Weng, Ee-Peng Lim, Jing Jiang
School of Information Systems
Singapore Management University
{jsweng,eplim,jingjiang}@smu.edu.sg

Qi He
College of Information Sciences and Technology
Pennsylvania State University
qhe@ist.psu.edu

ABSTRACT

This paper focuses on the problem of identifying influential users of micro-blogging services. *Twitter*, one of the most notable micro-blogging services, employs a social-networking model called "following", in which each user can choose who she wants to "follow" to receive *tweets* from without requiring the latter to give permission first. In a dataset prepared for this study, it is observed that (1) 72.4% of the users in *Twitter* follow more than 80% of their followers, and (2) 80.5% of the users have 80% of users they are following follow them back. Our study reveals that the presence of "reciprocity" can be explained by phenomenon of *homophily* [14]. Based on this finding, TwitterRank, an extension of PageRank algorithm, is proposed to measure the influence of users in *Twitter*. TwitterRank measures the influence taking both the topical similarity between users and the link structure into account. Experimental results show that TwitterRank outperforms the one *Twitter* currently uses and other related algorithms, including the original PageRank and Topic-sensitive PageRank.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval model*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*

General Terms

Algorithms, Design, Experimentation

Keywords

Twitter, influential, PageRank

1. INTRODUCTION

Micro-blogging is an emerging form of communication. It allows users to publish brief message updates, which can be submitted in many different channels, including the Web and text messaging service [1, 16]. One of the most notable micro-blogging services is *Twitter*[1]. It allows *twitterers* to publish *tweets* (with a limit of 140 characters)[2]. *Twitter* also provides the "social-networking" functionality.

Unlike other social network services that require users to grant friend links to other users befriending them, *Twitter* employs a social-networking model called "following", in which each *twitterer* is allowed to choose who she wants to follow without seeking any permission. Conversely, she may also be followed by others without granting permission first. In one instance of "following" relationship, the *twitterer* whose updates are being followed is called the "*friend*", while the one who is following is called the "*follower*".

Twitter has gained huge popularity since the first day that it was launched [16, 4]. It has also drawn increasing interests from research community. There is previous work [8] to study the topological and geographical properties of the social network formed by the *twitterers* and their followers. In this paper, we are interested in identifying the influential *twitterers*[3]. The benefit of solving this problem is multifold. First, it potentially brings order to the real-time web[4] in that it allows the search results to be sorted by the authority/influence of the contributing *twitterers* giving a timely update of the thoughts of influential twitterers. Second, according to [16], *Twitter* is also a marketing platform. Targeting those influential users will increase the efficiency of the marketing campaign [9, 10]. For example, a handphone manufacturer can engage those *twitterers* influential in topics about IT gadgets to potentially influence more people. There are also applications that utilize *Twitter* to gather opinions and information on particular topics. Identifying influential *twitterers* for interesting topics can improve the quality of the opinions gathered.

Currently, *Twitter* and many other applications interpret a *twitterer*'s influence as the number of followers she has. However, is this really a good indicator of influence? In a dataset prepared for this study, it is observed that (1) 72.4% of the users follow more than 80% of their *followers*, and (2) 80.5% of the user have 80% of their *friends* follow them back. Two seemingly conflicting reasons can possibly

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WSDM'10, February 4–6, 2010, New York City, New York, USA.
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...$10.00.

[1] Another similar service is *Plurk*.

[2] Users in *Twitter* are usually dubbed *twitterers*, and the short message updates published by the users *tweets*.

[3] In this paper, an influential *twitterer* is one with certain authority within her social network.

[4] Real-time web: http://en.wikipedia.org/wiki/Real-time_web.

explain such "reciprocity". First, the "following" relationship is so casual that each *twitterer* just randomly follows someone, and those being followed follow back just for the sake of courtesy. Second, it might be the opposite, i.e., the "following" relationship is a strong indicator of the similarity among users. In other words, a *twitterer* follows a friend because she is interested in the topics the friend publishes in *tweets*, and the friend follows back because she finds they share similar topic interest. This phenomenon is called "*homophily*", which has been observed in many social networks [14]. The cause of such "reciprocity" has important implication here. If it is caused by the first reason, identifying the influential *twitterers* based on "following" relationship would be rendered meaningless since the "following" relationship itself does not carry strong indication of influence. On the other hand, the presence of *homophily* indicates that the "following" relationships between *twitterers* are related to their topical similarity.

Our study confirms that *homophily* does exist in the context of *Twitter*. This justifies that there are some *twitterers* who do seriously "follow" someone because of common topical interests instead of just playing a "number game". Based on this observation, we propose a novel approach to measure the influence of *twitterer*s, known as TwitterRank. The framework of the proposed approach is shown in Figure 1. First, topics that *twitterers* are interested in are distilled automatically by analyzing the content of their *tweets*. Based on the topics distilled, topic-specific relationship networks among *twitterers* are constructed. Finally, we apply our TwitterRank algorithm, which is an extension of PageRank, to measure the influence taking both the topical similarity between *twitterers* and the link structure into account.

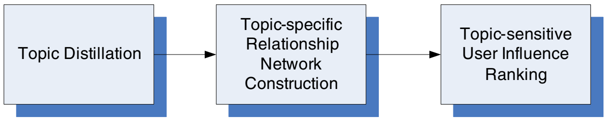

Topic Distillation
Topic-specific Relationship Network Construction
Topic-sensitive User Influence Ranking

Figure 1: Framework of the Proposed Approach

This paper improves the state-of-the-art by making two contributions. First, to the best of our knowledge, this paper is the first to report *homophily* in *Twitter*. Second, it introduces TwitterRank to measure the topic-sensitive influence of the *twitterers*. Prior to this, a *twitterer*'s influence is often measured by her node in-degree in the network, i.e., the number of *followers*. However, as observed in previous social network analysis studies [12, 3], in-degree does not accurately capture the notion of influence. PageRank improves over in-degree by considering the link structure of the whole network [3]. Nevertheless, Pagerank ignores the interests of *twitterers*, which affects the way *twitterers* influence one another. Our proposed approach addresses the shortcomings of in-degree and PageRank by taking into account both the link structure and topical similarity among *twitterers*.

The rest of this paper is organized as follows: A *Twitter* dataset has been prepared for the purpose of this study. Section 2 describes in detail how the dataset is prepared. Topic distillation and the phenomenon of *homephily* observed in the dataset is elaborated in Section 3, while TwitterRank is proposed in Section 4. Section 5 presents the experimental results, comparing TwitterRank with the benchmark method currently used by *Twitter* and other related algorithms. Section 6 briefly summarizes related work. Finally, Section 7 concludes with directions for further research.

2. TWITTER DATASET

For the purpose of this study, a set of *Twitter* data about Singapore-based *twitterers* was prepared in April, 2009 as follows:

1. We obtained a set of top-1000 Singapore-based *twitterers*[5] from *twitterholic.com*[6]. Denote this set as $\mathcal{S}$. As four of the top-1000 *twitterers* were not available when the dataset was being prepared, $|\mathcal{S}| = 996$.

2. We then crawled[7] all the *followers* and *friends* of each individual *twitterer* $s \in \mathcal{S}$ and stored them in set $\bar{\mathcal{S}}$.

3. Let $\mathcal{S}' = \mathcal{S} \bigcup \bar{\mathcal{S}}$, and $\mathcal{S}^* = \{s | s \in \mathcal{S}'$, and s is from Singapore$\}$. $|\mathcal{S}^*| = 6748$. For each $s \in \mathcal{S}^*$, we crawled[7] all the tweets she had published so far. Denote the set of all the tweets obtained as $\mathcal{T}$. $|\mathcal{T}| = 1,021,039$.

2.1 Tweet Distribution

The latest *tweet* in the dataset was published on April 25, 2009, while the earliest one was on July 18, 2006. Numbers of *tweets* by month during the time period captured in the dataset are plotted in Figure 2. It shows that *Twitter* started to attract substantial attention from Singapore-based *twitterers* from March 2008 onwards.

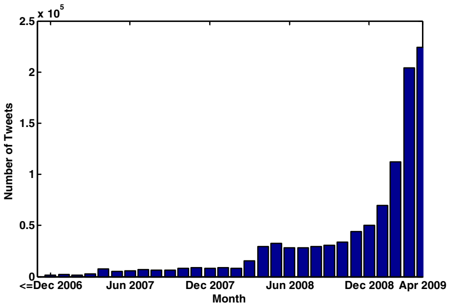


Figure 2: Number of *Tweets* per Month

Out of the 6748 *twitterers* in the dataset, only 5686 publish at least one *tweet*. For those 5686 *twitterers*, the average number of *tweets* each publishes is 179.57. The distribution of the *tweets* per *twitterer* is shown in Figure 3. If we do not consider the "outliers" indicated by the red circle[8], it follows a power-law distribution. The presence of "outliers" in the dataset is caused by a restriction implemented by *Twitter*, which limits the maximum number of *tweets* visible to be 3200 even a *twitterer* has published more than 3200 *tweets*.

There are 30 such active *twitterers* in the dataset. Four of them are bots that publish *tweets* directly obtained from

[5] Those with "location" specified as "Singapore" in their profiles are considered Singapore-based.

[6] Same as *Twitter*, *twitterholic* finds top *twitterers* based on the number of *followers*.

[7] It is noted that we do not use Twitter API for data crawling due to the default hourly API limit. Instead, we crawl and parse each *twitterer*'s profile page to obtain the *tweets* published, *followers* list, and *friends* list.

[8] It is noted that the single dot in the figure denotes the number of "outliers" but not one single such "outlier".

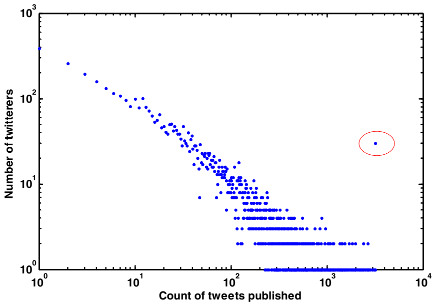
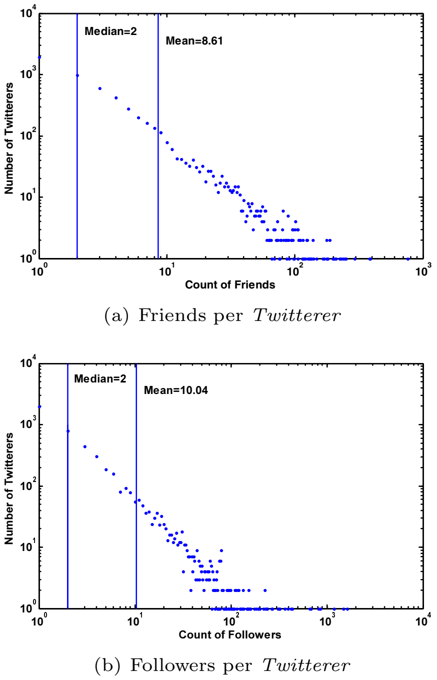


Figure 3: Distribution of *Tweets* per *Twitterer*

RSS feeds (usually more than one feed) they have subscribed to. We excluded two bots, one always re-publishing *followers*' tweets and another publishing only numbers. A spammer frequently publishing his username and URL in *tweets* was also excluded.



(a) Friends per *Twitterer*

(b) Followers per *Twitterer*

Figure 4: Distribution of the Following Relationships

2.2 Friends/Followers

There are in total 49872 "following" relationships among the *twitterers* in $\mathcal{S}^*$. Among the 6745 *twitterers*, 957 have no *friends*, while 1782 have no *followers*. The distribution of the numbers of the friends/followers each *twitterer* has are plotted in Figures 4(a) and 4(b) respectively. They again follow power-law distribution.

2.3 Reciprocity in Following Relationships

Reciprocity in *following* relationships is prevalent in *Twitter*. We examine this reciprocity by showing the correlation between number of *friends* and number of *followers* for each *twitterer* in Figure 5. It shows that the more *friends* a *twitterer* has, the more *followers* she has, and vice versa. A closer examination of the dataset reveals that there is high chance of "reciprocity" presented in the "following" relationships.

- 72.4% of the *twitterers* follow more than 80% of their followers,
- and 80.5% of the *twitterers* have 80% of their friends follow them back.

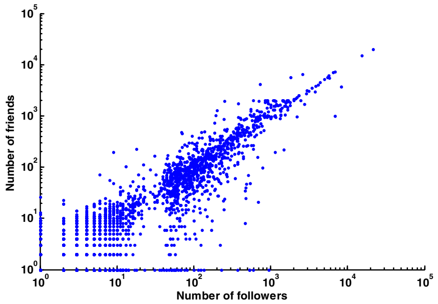


Figure 5: Number of Friends vs. Number of Followers

3. HOMOPHILY IN TWITTER

As mentioned in Section 1, two conflicting reasons can possibly explain such a "reciprocity", i.e., *twitterers*' casual "following" behaviors versus *homophily*. *Homophily* is a phenomenon showing that people's social networks "are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics" [14]. In the context of *Twitter*, *homophily* implies that a *twitterer* follows a *friend* because she is interested in some topics the *friend* is publishing, and the *friend* follows back because she finds they share similar topical interest.

Although it is beyond the scope of this paper to find the real cause of the "reciprocity" in the "following" relationships for each *twitterer*[9], the presence of *homophily* implies that there are *twitterers* who are serious in choosing *friends* to follow. This implication is important in that identifying the influential *twitterers* based on the "following" relationships would be rendered meaningless if no *twitterer* is serious in "following" others. Two questions would help to verify whether *homophily* presents in the context of *Twitter*:

Question 1: Are *twitterers* with "following" relationships more similar than those without according to the topics they are interested in?

Question 2: Are *twitterers* with reciprocal "following" relationships more similar than those without according to the topics they are interested in?

To answer these questions, we need to know the topics that *twitterers* are interested in and to measure the topical similarity between pairs of *twitterers*. However, topic interests are not explicitly expressed by *twitterers*. A possible

[9] In fact, our experimental results show that the two reasons may co-exist in the context of *Twitter*.

solution is to use the "#hashtag" uttered by *twitterers*[10]. Nevertheless, there is a very low usage of "#hashtag" in the dataset, which makes "#hashtag"s not appropriate to be used as topics. To overcome this challenge, topic modeling, which is commonly used to analyze large volumes of unlabeled contents, is applied to automatically distill topics.

3.1 Topic Distillation

The goal of the topic distillation is to automatically identify the topics that *twitterers* are interested in based on the *tweets* they published. For this purpose, **Latent Dirichlet Allocation (LDA)** model [2, 18, 6] is applied, which is an unsupervised machine learning technique to identify latent topic information from large document collection. It uses a "bag of words" assumption, which treats each document as a vector of word counts. Based on this assumption, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. It also assumes a generative process for generating each document as follows:

1. for each document, pick a topic from its distribution over topics,

2. sample a word from the distribution over the words associated with the chosen topic,

3. the process is repeated for all the words in the document.

More formally, each of a collection of D documents is associated with a multinominal distribution over T topics, which is denoted as θ. Each topic is associated with a multinomial distribution over words, denoted as ϕ. θ and ϕ have Dirichlet prior with hyper-parameters α and β respectively. For each word in one document d, a topic z is sampled from the multinomial distribution θ associated with the document, and a word w from the multinomial distribution ϕ associated with topic z is sampled consequently. This generative process is repeated N_d times (N_d is the total number of words in document d) to form document d [2, 18, 6]. This generative process can be graphically represented using commonly-used plate notation in Figure 6. In this figure, shaded and unshaded plates indicate observed and latent variables respectively. An arrow corresponds to a conditional dependency between two variables and boxes indicate repeated sampling with the number of repetitions given by the variable in the bottom of the corresponding box.

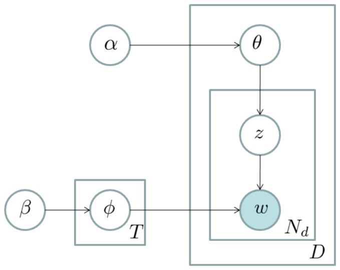


Figure 6: Graphical Representation of LDA Model

The model has two parameters to be inferred from the data, i.e. document-topic distributions θ, and the T topic-word distributions ϕ. By learning these two parameters, information can be obtained about which topics authors typically write about as well as a representation of the content of each document in terms of these topics. In this study, Gibbs sampling is applied for model parameter estimation[11].

To distill the topics that *twitterers* are interested in using LDA, documents should naturally correspond to *tweets*. However, since the goal is to understand the topics that each *twitter* is interested in rather than the topic that each single *tweet* is about, we aggregate the *tweets* published by individual *twitterer* into a big document. Thus, each document essentially corresponds to a *twitterer*.

The result is represented in three matrices:

1. DT, a $D \times T$ matrix, where D is the number of *twitterers* and T is the number of topics. DT_{ij} contains the number of times a word in *twitterer* s_i's *tweets* has been assigned to topic t_j.

2. WT, a $W \times T$ matrix, where W is the number of unique words used in the *tweets* and T is the number of topics. WT_{ij} captures the number of times unique word w_i has been assigned to topic t_j,

3. and Z, a $1 \times N$ vector, where N is the total number of words in the *tweets*. Z_i is the topic assignment for word w_i.

3.2 Hypothesis Testing

Among the three matrices in the result of topic distillation, matrix *DT* is of particular interest. It contains the number of times a word in a *twitterer*'s *tweets* has been assigned to a particular topic. We can row normalize it as DT' such that $\|DT'_{i\cdot}\|_1 = 1$ for each row $DT'_{i\cdot}$. Each row of matrix DT' is basically the probability distribution of *twitterer* s_i's interest over the T topics, i.e. each element DT'_{ij} captures the probability that *twitterer* s_i is interested in topic t_j. Given this, the topical difference between *twitterers* can be measured as follows.

Definition 1. *Topical difference between two* twitterers *s_i and s_j can be calculated as:*

$$dist(i,j) = \sqrt{2 * D_{JS}(i,j)}\text{[12]}. \tag{1}$$

$D_{JS}(i,j)$ is the Jensen-Shannon Divergence between the two probability distributions $DT'_{i\cdot}$ and $DT'_{j\cdot}$, which is defined as:

$$D_{JS}(i,j) = \frac{1}{2}(D_{KL}(DT'_{i\cdot}\|M) + D_{KL}(DT'_{j\cdot}\|M)) \tag{2}$$

M is the average of the two probability distributions, i.e. $M = \frac{1}{2}(DT'_{i\cdot} + DT'_{j\cdot})$. D_{KL} in Eq (2) is the Kullback-Leibler Divergence which defines the divergence from distribution Q to P as: $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$. ∎

[10] *Twitter* allows each *twitterer* to associate a hashtag to the *tweets*. Hashtags in *tweets* are equivalent to tags typically found in content sharing services, e.g. *Youtube* and *Flickr*.

[11] Conventional Gibbs sampling is applied instead of collapsed Gibbs sampling, though collapsed Gibbs sampling is shown to achieve faster convergence and better performance [17].

[12] It has been proved $\sqrt{2 * D_{JS}(i,j)}$ is a metric for probability distributions which fulfills the triangle inequality [5]. Another reason it is used here is that it reduces the non-normality of the data, which will potentially influence the robustness of the t-test.

With the topical difference measured, the two questions listed in the beginning of Section 3 can be answered by statistical hypothesis testing. It is noted that, in this study, hypothesis testing, and topic distillation as well, is applied on a set of *twitterers* who publish more than 10 *tweets* in total. We denote this set as $\mathcal{S}_u^*$, and $|\mathcal{S}_u^*| = 4050$.

3.2.1 Question 1

Question 1 can be formalized as a two-sample t-test:

Let μ_{follow} be the mean topical difference of the pairs of *twitterers* with "following" relationship, and $\mu_{nofollow}$ the mean topical difference of those without.

The null hypothesis is $H_0 : \mu_{follow} = \mu_{nofollow}$, and the alternative hypothesis is $H_1 : \mu_{follow} < \mu_{nofollow}$.

Ideally, individual statistical hypothesis testing shall be conducted for each *twitterer*. Nevertheless, most of the *twitterers* (3785 out of 4050) have less than 30 *friends*, which is not statistically significant. Therefore, two cases are considered when answering **Question 1**.

Case 1.

Denote the set of *twitterers* with more than 30 friends as $\mathcal{S}_{U_a}^*$, and $|\mathcal{S}_{U_a}^*| = 265$. Individual statistical hypothesis test is conducted for every *twitterer* $s_i \in \mathcal{S}_{U_a}^*$. First, calculate the topical difference between s_i and each of her friends, based on which μ_{follow} is calculated. Then, choose some *twitterers* uniformly at random from those s_i does not follow, and the number of the chosen *non-friend*s is same as the number of s_i's friends. Calculate the topical difference between s_i and each *non-friend*, based on which $\mu_{nofollow}$ is calculated. Finally, a two-sample t-test (under the assumption of unequal population variances) is conducted on the two populations formed with the above approach.

Results shows that for 232 out of the 265 *twitterers* with more than 30 friends, the null hypothesis is rejected at significant level $\alpha = 0.01$[13].

Case 2.

Denote the set of *twitterers* with less than 30 friends as $\mathcal{S}_{U_b}^*$, and $|\mathcal{S}_{U_b}^*| = 3785$. For this set of *twitterers*, the hypothesis testing is conducted on the *twitterer* congregation. First of all, calculate the topical difference for all the pairs of *twitterers* whose "following" relationships are initiated by any *twitterer* $s_i \in \mathcal{S}_{U_b}^*$, based on which μ_{follow} is calculated. Then, for each $s_i \in \mathcal{S}_{U_b}^*$, choose some *non-friend*s uniformly at random, the number of the chosen *non-friend*s is same as the number of s_i's friends. Congregate all the pairs of *twitterers*, and calculate the difference between each pair, based on which $\mu_{nofollow}$ is calculated. Finally, a two-sample t-test (under the assumption of unequal population variances) is conducted on the two populations formed with the above approach. The test outcome is that the null hypothesis is rejected at significant level $\alpha = 0.01$ with a p-value of $4.5 * 10^{-6}$.

Together with the results in **Case 1**, the answer to **Question 1** is clearly that with very high probability, *twitterers* with "following" relationships are more similar than those without according to the topics they are interested in.

3.2.2 Question 2

Question 2 is also formalized as a two-sample t-test:

Let μ_{sym} be the mean topical difference of the pairs of *twitterers* with reciprocal "following" relationship, and μ_{asym} the mean topical difference of those without.

The null hypothesis is $H_0 : \mu_{sym} = \mu_{asym}$, and the alternative hypothesis is $H_1 : \mu_{sym} < \mu_{asym}$.

There are in total 11505 pairs of *twitterers* with reciprocal "following" relationship. However, there are only 67 *twitterers* with more than 30 reciprocal and non-reciprocal friends respectively. Hence, we conduct the two-sample t-test on the *twitterer* congregation. First of all, calculate the topical difference for all the pairs of *twitterers* with reciprocal "following" relationship, based on which μ_{sym} is calculated. Then, for each *twitterer*, choose some non-reciprocal *friends* uniformly at random such that the number of the chosen non-reciprocal *friends* is same as the number of reciprocal *friends* she has. Congregate all the non-reciprocal relationships, and calculate the topical difference for each non-reciprocal relationship, based on which μ_{asym} is calculated. With the above two populations, the null hypothesis is rejected at significant level $\alpha = 0.01$ with a p-value of $1.2 * 10^{-6}$. This outcome gives a positive answer to **Question 2** that with very high probability, *twitterers* with reciprocal "following" relationships are more similar than those without according to the topics they are interested in.

Positive answers to both **Question 1** and **Question 2** provide evidences to the existence of the *homophily* phenomenon in the *Twitter* dataset. Based on this finding, a novel approach to measure *twitterers*' influence is proposed in the next section.

4. TOPIC-SENSITIVE INFLUENCE MEASURE

Intuitively, the influence of a *twitterer* can be interpreted similar to the "authority" of a web page: a *twitterer* has high influence if the sum of influence of her *followers* is high; at the same time, her influence on each *follower* is determined by the relative amount of content the follower received from her. This similarity motivates the use of PageRank in measuring influence.

Although the "authority" of web page and influence of *twitterer* shares certain similarities, there are also major differences. The *influence* on each follower is purely based on relative amount of content the *follower* receives as the latter may not read content with topics less interesting even when the relative content is large. Since *twitterers* generally have different expertise and/or interests in various topics, influence of *twitterers* also vary in different topics. Given this, a topic-sensitive *TwitterRank* is proposed to measure the influence of *twitterers*.

4.1 Topic-specific TwitterRank

First of all, a directed graph $D(V, E)$ is formed with the *twitterers* and the "following" relationships among them. V is the vertex set, which contains all the *twitterers*. E is the edge set. There is an edge between two *twitterers* if there is "following" relationship between them, and the edge is directed from *follower* to *friend*.

[13] The robustness of t-test depends very much on the normality of the population. When the population is not normal, the skewness and kurtosis of the distribution will affect the value of the test statistic t. Nevertheless, according to [15], for skewness and kurtosis in the range of $[0, 0.7]$ and $[-0.5, 4]$ respectively, their effort on the t value is small, in which case the t-test result is still valid. Skewness and kurtosis of the data in all the t-tests in this paper are in these ranges.

A random surfer model on graph D computes the TwitterRank as follows: the random surfer visits each *twitterer* with certain probability by following the appropriate edge in D. TwitterRank differentiates itself from PageRank in that the random surfer performs a topic-specific random walk, i.e. the transition probability from one *twitterer* to another is topic-specific. By doing so, we are essentially constructing a topic-specific relationship network among *twitterers*.

The transition matrix for topic t, denoted as P_t, is defined as follows.

Definition 2. *Given a topic t, each element of matrix P_t, i.e. the transition probability of the random surfer from* follower *s_i to* friend *s_j, is defined as:*

$$P_t(i,j) = \frac{|\mathcal{T}_j|}{\sum_{a:\ s_i\ follows\ s_a} |\mathcal{T}_a|} * sim_t(i,j) \tag{3}$$

$|\mathcal{T}_j|$ is number of tweets *published by s_j, and $\sum_{a:\ s_i\ follows\ s_a} |\mathcal{T}_a|$ sums up the number of* tweets *published by all of s_i's* friends. *$sim_t(i,j)$ in Eq. (3) is the similarity between s_i and s_j in topic t, which is defined as:*

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}| \tag{4}$$

∎

This definition captures two notions. Assume *twitterer* s_i follows a number of *friends*. Those *friends* publish different numbers of *tweets*, all of which will be directly visible to s_i. The more a *friend* s_j publishes, the higher portion of *tweets* s_i reads is from s_j. Generally, this leads to a higher influence on s_i, which corresponds to a higher transition probability from s_i to s_j. This intuition is captured in the first term in the RHS of Eq. (3). Figure 7 shows an example about three *twitterers*. s_c follows s_a and s_b, who publish 500 and 1000 *tweets* respectively. In this case, s_b's influence on s_c is two times of that of s_a, when the topical similarity among the three *twitterers* is not taken into account.

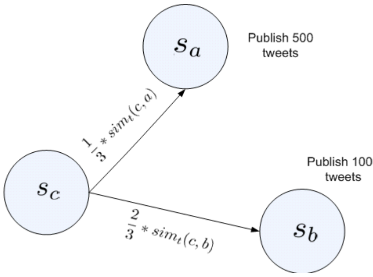


Figure 7: Example of Transition Probability Calculation

Second, s_j's influence on s_i is also related to the topical similarity between the two as suggested by the *homophily* phenomenon discussed in Section 3. Row-normalized matrix DT' is one of the results in the topic distillation. A row $DT'_{j\cdot}$ contains the probability of *twitterer* s_j's interest in different topics. The similarity between s_i and s_j in topic t can be evaluated as the difference between the probability that the two *twitterers* are interested in the same topic t, which is basically the second term in the RHS of Eq. (3). The more similar the two *twitterers* are, the higher the transition probability from s_i to s_j.

It is possible that some *twitterers* would "follow" one another in a looping manner without "following" other *twitterers* outside the loop. Such loop will accumulate high influence without distribute their influence. To tackle this, a teleportation vector E_t is also introduced, which basically captures the probability that the random surfer would "jump" to some *twitterers* instead of following the edges of the graph D. E_t is defined as follows.

Definition 3. *The teleportation vector of the random surfer in topic t is defined as:*

$$E_t = DT''_{\cdot t} \tag{5}$$

$DT''_{\cdot t}$ is the t-th column of matrix DT'', which is the column-normalized form of matrix DT such that $\|DT''_{\cdot t}\|_1 = 1$. DT is one of the results obtained during the topic distillation, each entry of which contains the numbers of times words in a twitterer*'s tweets has been assigned to a specific topic.* ∎

With the transition probability matrix and teleportation vector defined, the topic-specific *TwitterRank* can be calculated.

Definition 4. *The topic-specific* TwitterRank *of the* twitterers *in topic t, denoted as $\overrightarrow{TR_t}$, can be calculated iteratively by:*

$$\overrightarrow{TR_t} = \gamma P_t \times \overrightarrow{TR_t} + (1-\gamma)E_t \tag{6}$$

P_t is the transition probability matrix defined in Eq. (3), E_t is the teleportation vector defined in Eq. (5). γ is a parameter between 0 and 1 to control the probability of teleportation. The lower γ is, the higher probability the random surfer will teleport to twitterers *according to E_t, and vice versa.* ∎

4.2 Aggregation of Topic-specific TwitterRank

The approach presented in Section 4.1 generates a set of topic-specific TwitterRank vectors, which basically measure the *twitterers*' *influence* in individual topics. An aggregation of *TwitterRank* can also be obtained to measure *twitterers*' overall *influence*.

Definition 5. Twitterers*'s general* influence *can be measured as an aggregation of the topic-specific TwitterRank in different topics, which is calculated as:*

$$\overrightarrow{TR} = \sum_t r_t \cdot \overrightarrow{TR_t} \tag{7}$$

$\overrightarrow{TR_t}$ is the TwitterRank vector for topic t, while r_t is the weight assigned to topic t and associated $\overrightarrow{TR_t}$. ∎

Depending on the applications, different set of weights can be assigned to derive the influence of *twitterers* in different scenarios.

General influence: r_t's can be set as the probabilities of different topics' presence, which are calculated according to the number of times unique words have been assigned to corresponding topics as captured in matrix WT. In this case, the aggregation of TwitterRank is essentially the *twitterers*' *general influence*.

Perceived general influence: r_t's can also be set as the probabilities that a particular *twitterer* s_i is interested in different topics, which are calculated according to the number of times words in s_i's *tweets* have been assigned to corresponding topics as captured in matrix DT. In this case, the aggregation of TwitterRank becomes s_i's personal perception of *twitterer*s' *general influence*.

5. EMPIRICAL EVALUATION

Section 4 proposes TwitterRank, which measures different *twitterer*s' influence by taking into account the topical similarity among *twitterers* as well as the link structure. This section shows the results of applying TwitterRank in our *Twitter* dataset. We also elaborate on an evaluation procedure for effective comparison with other related algorithms.

5.1 Influential twitterers identified in the Twitter dataset

We first compare the most influential *twitterers* identified by TwitterRank with the most active *twitterers* identified during topic distillation.

As mentioned in Section 3.2, topic distillation is applied on a set of *twitterers* who publish more than 10 *tweets* in total. We denote this set as $\mathcal{S}_u^*$, and $|\mathcal{S}_u^*| = 4050$. All the experiments in the rest of this paper is conducted on this set of *twitterers* and their *tweets*. The *tweets* in the dataset are written with a mixture of different languages including Chinese, English, French, German, Japanese, etc. We removed from *tweets* those words containing non-English characters, stopwords, punctuations, numbers, URLs, words with less than 3 characters, and words in the form "@username". These words do not help in topic modeling. The remaining words are stemmed. LDA is conditioned on three parameters, i.e. Dirichlet hyper-parameters α, β, and topic number T. In this paper, they are set as $T = 50$, $\alpha = 50/T$, and $\beta = 0.1$[14]. Teleportation parameter in TwitterRank, i.e. γ in Eq. (6), is set as $\gamma = 0.85$.

Table 1 lists the top-5 active and influential *twitterers* in the five top topics. Top topics are identified in the order of the probabilities of topic presence, which are calculated according to the number of times unique words have been assigned to corresponding topics as captured in matrix WT (see Section 3.1). It is observed that the active *twitterers* are not necessarily influential in each topic.

The results in Table 1 are reasonable. *Twitterer*s "mrbrown", "moby74", and "kormmandos" are among the top-5 influential *twitterer*s in all the five top topics identified in the dataset. "mrbrown" mainly tweets about Singapore citizen life and IT-related news. He also tweets often about things happened in his office or during his trips, as well as those in his bicycle ride from home to office. The words frequently used in expressing these topics are captured in the five top topics as shown in Tables 1. Additionally, "mrbrown" has the highest number of *followers* (as captured in the dataset), including some influential ones like "AngMoGirl", "claudia10", and "moby74".

"moby74" tweets mainly about work, family life, food, and IT-related topics (such as the features of Twitter, website design, and Internet connection speed). Although "moby74" has much fewer *followers* than "mrbrown", "moby74" has many influential *followers* including "AngMoGirl", "claudia10", and "mrbrown".

"kormmandos" tweets frequently about food and blogging. He is also followed by a number of influential *twitterers*, including "AngMoGirl", "moby74" and "mrbrown".

"singaporenews" is identified among the top-5 influential *twitterers* in topic #1, #2, and #3. He is followed by a number of influential *twitterers* including "mrbrown". As the name suggests, he tweets mostly about news events in Singapore. Quite often, he also tweets about world news. In contrast, he tweets less about IT news and food-related topics, so he is not identified as the most influential ones in topic #4 and #5. Similar as "singaporenews", "sginfomap" tweets mostly about Singapore news. Nevertheless, he tweets about world news less frequently than "singaporenews". Both "singaporenews" and "sginfomap" are followed by "mrbrown".

"AngMoGirl" created her *Twitter* account in Feb 2009. During her early exploration of *Twitter*, "AngMoGirl" tweets frequently about the functionality of *Twitter*. She also has a number of influential *followers*, including "benkoe", "claudia10", "kormmandos", and "mrbrown".

"claudia10" is a active Singapore-based blogger. In both blog and *Twitter*, she mostly writes about IT gadgets, life style, and social media. She has a number of influential *followers*, including "AngMoGirl", "benkoe", "mrbrown", and "moby74".

"hana77" is a Singapore-based DJ, and often tweets about hair styles and her haircut appointments. She is followed by some *followers* who are followed by influential *twitterers*. For example, "hana77" is followed by "FashionlyNews", who is followed "AngMoGirl"; she is also followed by "slightlyfamous", who is followed by "claudia10".

"benkoe" is an employee associated with a Singapore-based social media analysis company, who often tweets about social media and IT gadgets. Often times, he uses words like "love" and "feel" to express his personal feeling about the subjects that he tweets. He is followed by some influential *twitterers*, including "AngMoGirl", "claudia10", "moby74", and "mrbrown".

5.2 Comparison with related algorithms

In this section, we study quantitatively the effectiveness of the proposed TwitterRank. Comparisons against related algorithms are also conducted. The related algorithms studied include:

- **In-degree**, which measures the influence of *twitterer*s by the number of *followers*. This is the measurement currently employed by *Twitter* and many other third-party services, such as *twitterholic.com* and *wefollow.com*.
- **PageRank**, which measures the influence with only link structure of the network taken into account [3].
- **Topic-sensitive PageRank**, which measures topic-specific influence by calculating PageRank vector for each topic. Nevertheless, unlike TwitterRank, same relationship network, i.e., same transition probability matrix is used for different topics, but with a topic-biased teleportation vector [7].

For ease of presentation, the proposed TwitterRank is denoted as **TR**, and the three related algorithms are abbreviated to **InD**, **PR**, and **TSPR** respectively.

[14] The choice of different values for these parameters has implications for the results of the model. This is basically a model selection problem. Nevertheless it is not discussed since the focus of this paper is how to identify the influential *twitterers* in different topics identified.

Table 1: Active and Influential Twitterers in Top Topics

Topic #	Associated Words	Active *Twitterers*	Influential *Twitterers*
1	work morn time night home	nikipaniki annsherry cblake slightlyfamous mintea	mrbrown moby74 kormmandos singaporenews sginfomap
2	people world life word time	ennn PatchouliW balaji_dutt PoonPiPi FunkeeMonk	mrbrown moby74 kormmandos singaporenews AngMoGirl
3	time twitter hope work friend	maynaseric asheraw stuarttan balaji_dutt derrickkwa	mrbrown moby74 singaporenews sginfomap kormmandos
4	googl design twitter web site	balaji_dutt BoltClock fabrikade flashmech erwanmace	mrbrown moby74 kormmandos claudia10 AngMoGirl
5	love feel eat hot hair	highpriestess tstar nikipaniki killerpussy moby74	mrbrown moby74 kormmandos hana77 benkoe

5.2.1 *Correlation*

We first study the correlation between the rank lists generated by the different algorithms. The correlation is measured as the Kendall's τ [11]. τ takes value in the range of $[-1, 1]$. If the two lists are exactly the same, $\tau = 1$; whereas $\tau = -1$ if one list is the reverse of the other. For other values in the range, a larger value of τ implies higher agreement between the two lists.

Table 2(a) lists the τ values between the rank lists generated by various algorithms studied. For **TR** and **TSPR**, we apply the across-topic aggregation mentioned as "*general influence*" in Section 4.2. It is observed that **TR** generates ranked list different from those generated by other algorithms since $\tau \neq 1$. It is also observed that **TR** has higher agreement with **TSPR** than with **InD** and **PR**. This is because both **TR** and **TSPR** consider the topical dimension while **InD** and **PR** do not. We have also studied the correlation between the four algorithms in different topics, the same trend is observed. Table 2(b) lists τ values between the rank lists by the four algorithms in the 5 top topics listed in Tables 1.

Table 2: Correlation between Rank Lists by Different Algorithms

(a) General Rank

	Kendall τ
TR vs. **InD**	0.4234
TR vs. **PR**	0.4719
TR vs. **TSPR**	0.6800

(b) Topic-specific Rank

	#1	#2	#3	#4	#5
TR vs. **InD**	0.4493	0.42	0.453	0.479	0.4025
TR vs. **PR**	0.4924	0.4581	0.5195	0.5111	0.4417
TR vs. **TSPR**	0.6902	0.6933	0.6815	0.6961	0.6944

5.2.2 *Performance in Recommendation Task*

Tangible benefit can be realized when applying it to some tasks. In this paper, we evaluate the usefulness of TwitterRank in the *twitterer* recommendation task. The recommendation task is designed as Figure 8 shows.

```
1 randomly choose $|L|$ existing "following" relationship
  formed among twitterers in $\mathcal{S}_u^*$;
2 foreach $l \in L$ do
3     let $s_o$ and $s_f$ be the follower and friend in
      "following" relationship $l$ respectively;
4     randomly choose 10 twitterers that $s_o$ does not
      follow, denote this set as $\mathcal{S}_t$;
5     remove $l$ to generate a new network in which twitter
      $s_o$ does not follow $s_f$;
6     apply different algorithms to measure the influence
      of $s_f$ and all the twitterers in $\mathcal{S}_t$ in the new network,
      based on which $s_o$ is recommended whether to
      "follow" $s_f$;
7     compare the quality of the recommendation by
      different algorithms;
8 end
```

Figure 8: Recommendation Task for Performance Evaluation and Comparison

L, the set of existing "following" relationships in Step 1 of the recommendation task is considered the "ground truth" for evaluation: the recommendation is considered "good" if s_f is ranked higher than all the *twitterers* in $\mathcal{S}_t$ chosen in Step 4. Given this, the quality of the recommendation is measured as the number of *twitterers* in $\mathcal{S}_t$ who have a higher rank than s_f. More formally, it is defined as follows:

Definition 6. *Assume l is a ranked list recommended by any of the algorithms, and s_i is a* twitterer. *Let $l(s_i)$ be the rank of s_i in l (a higher rank corresponds to a low-numbered rank in l). The quality of the recommendation $Q(l)$ is measured as $Q(l) = |\{s_i | s_i \in \mathcal{S}_t,\ and\ l(s_i) < l(s_f)\}|$. s_f is the* friend *removed in Step 5 in Figure 8. The lower the value of $Q(l)$ is, the higher the quality of corresponding algorithm is.* ∎

Different L's based on various criteria have been used to study the proposed TwitterRank's performance as comprehensively as possible. Currently, there are in total four criteria based on which L is generated:

(a): Two L's denoted by L_{fh} and L_{fl} are generated based on the number of *followers* that s_f has: L_{fh} has s_f with high *follower* count, while L_{fl} has s_f with low *follower* count. s_f's *follower* count is considered high if it is larger than FH, and low if smaller than FL. FH and FL are set as the 90*th* and 10*th* percentile of all the *follower* counts of the *twitterers* in $\mathcal{S}_u^*$. To generate L_{fh} (or L_{fl}), $|L| = 30$ "following" relationships are chosen uniformly at random among all the existing relationships in which s_f fulfills the criteria described above.

(b): Two L's denoted by L_{th} and L_{tl} are generated based on the number of *tweets* that s_f has. These two sets are generated in a similar approach as in (a). The difference is that the thresholds for high *tweet* count and low *tweet* count, denoted as TH and TL, are set as the $90th$ and $10th$ percentile of all the *tweet* counts of the *twitterers*.

(c): Two L's denoted by L_{dl} and L_{dh} are generated based on the topical difference between s_o and s_f. These two sets are generated in a similar approach as in (a) and (b). The difference is that the thresholds for low topical difference and high topical difference, denoted as DL and DH, are set as the $10th$ and $90th$ percentile of the difference of all the existing "following" relationships. The topical difference of a "following" relationship is measured according to Definition 1.

(d): Two L's denoted by L_{rr} and L_{ur} are generated based on whether there is reciprocal "following" relationship between s_o and s_f. There is no threshold applied. $|L| = 30$ "following" relationships are chosen uniformly at random among all the existing reciprocal relationships to generate a set L_{rr}, while L_{ur} is generated by randomly choosing 30 unilateral relationships.

There are eight sets of L used in each individual round of evaluation. Five rounds of evaluation are conducted.

Figure 9 shows the average results of the four algorithms with different sets of L over all the evaluation rounds. It is noted that, the ranked lists recommended by **TR** and **TSPR** are an aggregation of lists in different topics, and "perceived general influence" (see Section 4.2) is applied for aggregation. This aggregation scheme is more appropriate as it reflects the perception of the *twitterer* following others. It can be observed that all the algorithms perform better in scenarios where L_{dl} is used than in those where L_{dh} is used. This observation shows that there are *twitterers* who "follow" because of the topical similarity between them and their *friends*. This supports the phenomenon of *homophily* discussed in Section 3. From Figure 9, it is also observed that although **TR** does not outperform the other algorithms consistently, it achieves the best recommendation quality in most of the scenarios.

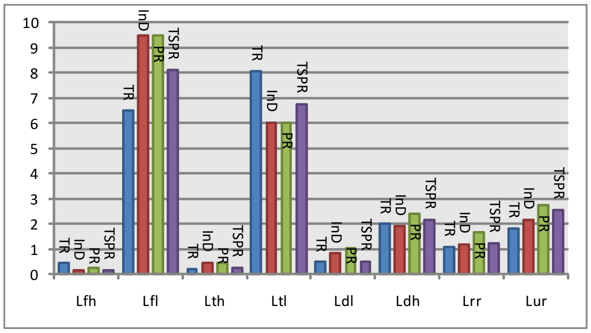


Figure 9: Comparison of Performance (measured by $Q(l)$) in the Recommendation Task

TR is outperformed by other algorithms in 3 out of the 8 scenarios studied, including those where L_{fh}, L_{tl}, and L_{dh} are used. In scenarios where L_{fh} is used, there is no obvious difference in the performance of all the algorithms. Yet, **InD** achieves the best performance. This is probably because, in the dataset, *twitterers*' "following" behaviors have already been biased toward those with more *followers*, since **InD** is essentially the algorithm applied in *Twitter* to recommend *friends*.

In scenarios where L_{tl} is used, **TR**'s performance is the worst among all. This is because the quality of topics distilled for s_f is not as good since LDA-based topic distillation is less accurate with little content available. Consequently, this impacts the performance of **TR** which takes into account the topical similarity when measuring the *twitterers*' influence.

In scenarios where L_{dh} is used, **TR** outperforms all the other algorithms except **InD**. This phenomenon, together with the one observed in scenarios where L_{fh} is used, shows that there still exist some *twitterers* who do not "follow" based on topical similarity, although *homophily* is observed.

TR performs the best in all the other scenarios, though the improvement is not significant in most of cases. It is noted that in scenarios where L_{fl} is used, **TR** outperforms the other algorithms significantly, especially **InD** and **PR**. This is because *friends* of s_f in the "following" relationships in L_{fl} are with lower numbers of *followers*. Consequently, the corresponding s_o would have lower chance to be biased by the recommendation made by *Twitter*, which is essentially made with **InD**. In such cases, the chance that the "following" relationship is formed due to topical similarity is higher. Therefore, **TR** outperforms **InD** and **PR**, which do not take into account topical similarity. Furthermore, **TR** outperforms **TSPR**. This is because **TSPR** uses identical transition probability matrix when calculating the topic-specific ranks. By doing so, **TSPR** basically propagates a *twitterer*'s influence in one topic to her *friends* in different topics with equal probabilities.

6. RELATED WORK

Currently, *Twitter* measures a *twitterer*'s influence as the number of *followers* she has. The more *followers* she has, the more impact she appears to make in the *Twitter* context, because she seems more popular. The underlying assumption here is that every *tweet* published by a *twitterer* is read by all her *followers*. A similar metric relies on the ratio between the number of one's *followers* and the number of *friends*. Another metric proposed by the Web Ecology project [13] measures the influence based on the ratio of attention (including *retweet*, *reply*, and *mention*) a *twitterer* received to the *tweets* she published.

These three metrics do not utilize the global link structure among *twitterers*. There are attempts which take into account the global link structure when measuring influence in the *Twitter* context, e.g. TunkRank[15]. TunkRank extends PageRank and calculates the influence of *twitterer* recursively as:

$$Influence(X) = \sum_{Y \in Followers(X)} \frac{1+p*Influence(Y)}{|Friends(Y)|}.$$

Here, p is the constant probability that *twitterers* retweet a *tweet*. TunkRank measures *twitterer* X's influence as the expected number of *twitterers* who will read a *tweet* that

[15] TunkRank is originally proposed by Daniel Tunkelang in http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/. An implementation of the idea is available at http://tunkrank.com/.

she publishes.In this respect, TunkRank is similar to the proposed TwitterRank (see the first term in the RHS of Eq. (3)). However, TunkRank (and the above-mentioned three metrics as well) ignores the possibility for *twitterers* to interact with the content in *Twitter*.

The proposed TwitterRank acknowledges such possibility and extends PageRank with the consideration of topical similarity between *twitterers*. The most similar work in this aspect is Topic-sensitive PageRank (**TSPR**) proposed by Haveliwala [7]. It is also this work that the performance of TwitterRank is compared against. **TSPR** uses identical transition probability matrix when calculating the topic-specific influence. By doing so, **TSPR** basically propagates a *twitterer*'s influence in one topic to her *friends* in different topics with equal probabilities. In contrast, TwitterRank applies different transition probability matrices for different topics, which is validated by the experimental results to capture the topic-specific influence better.

7. CONCLUSIONS AND FUTURE WORK

This paper focuses on finding influential *twitterers* in *Twitter*. This paper is the first to report the phenomenon of *homophily* in a community of *Twitter*. By making use of this phenomenon, a Pagerank-like algorithm, called TwitterRank, is proposed to measure the topic-sensitive influence of the *twitterers*. The experimental results shows that the proposed TwitterRank outperforms other related algorithms. Nevertheless, as an early attempt to bring order to *Twitter*, TwitterRank still has space for improvement.

First, as the experimental results show, there are still some *twitterers* "follow" not because of the topical similarity between them and their *friends*. We plan to classify different categories of *twitterers* by studying their "following" behaviors more closely, and apply TwitterRank on those with more serious "following" behaviors. Second, the current design of TwitterRank takes into account number of *tweets* a *twitterer* publishes (see Eq. (3)). This makes it susceptible to manipulations if a *twitterer* deliberately publishes a large number of *tweets*. In the future, we plan to improve this by incorporating other interactions between two *twitterers*, e.g. *reply/mention* between two *twitterers*. Third, we also plan to valid the "homophily" phenomenon and TwitterRank in a larger dataset. To collect a larger dataset, we are currently monitoring the *Twitter* public timeline using *Twitter* streaming API[16]. At the same time, we crawl the "following" relationship among *twitterers* using *Twitter* API[17]. Last but not least, currently the topic distillation is conducted on a snapshot of *Twitter* and the numbers of *twitterers* and topics are fixed. Nevertheless, *Twitter* is a platform for free and open conversations among *twitterers*. An incremental approach to topic distillation in *Twitter* is still a topic deserves further study.

8. REFERENCES

[1] Micro-blogging. http://en.wikipedia.org/wiki/Micro-blogging.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Network and ISDN Systems*, 30(1-7):107–117, 1998.

[4] A. Cheng and M. Evans. Inside Twitter: An in-depth look inside the Twitter world. http://www.sysomos.com/insidetwitter/, June 2009.

[5] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[7] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[8] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[9] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[10] D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP 2005: Proceedings of the 32nd International Colloquium on Automata, Languages and Programming*, pages 1127–1138, 2005.

[11] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[13] A. Leavitt, with Evan Burchard, D. Fisher, and S. Gilbert. The influentials: New approaches for analyzing influence on twitter. a publication of the Web Ecology project. http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf, Sept 2009.

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

[15] R. G. Miller. *Beyond ANOVA, basics of applied statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1986.

[16] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. Twitter and the micro-messaging revolution: Communication, connections, and immediacy–140 characters at a time. O'Reilly Report, November 2008.

[17] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577, New York, NY, USA, 2008. ACM.

[18] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, In Press.

[16] Streaming API: http://apiwiki.twitter.com/Streaming-API-Documentation.

[17] *Twitter* API: http://apiwiki.twitter.com/Twitter-API-Documentation.