



A Data-Based Approach to Discovering Multi-Topic Influential Leaders

Xing Tang, Qiguang Miao*, Shangshang Yu, Yining Quan

School of Computer Science and Technology, Xidian University, Xi'an, China

* ggmiao@126.com



OPEN ACCESS

Citation: Tang X, Miao Q, Yu S, Quan Y (2016) A Data-Based Approach to Discovering Multi-Topic Influential Leaders. PLoS ONE 11(7): e0158855. doi:10.1371/journal.pone.0158855

Editor: Wen-Bo Du, Beihang University, CHINA

Received: April 18, 2016
Accepted: June 22, 2016
Published: July 14, 2016

Copyright: © 2016 Tang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The Weibo dataset used in this paper are mining from Weibo.com(www.weibo.com). We confirm that the collection method complies with the term of service. To agree with weibo's terms of service, we cannot host the data on an external website. However, the data are available on request from the author(email:qqmiao@126.com).

Funding: The work was jointly supported by the National Natural Science Foundations of China under grant No. 61472302, 61272280, U1404620, and 41271447; The Open Projects Program of National Laboratory of Pattern Recognition (201600031); The Program for New Century Excellent Talents in University under grant No. NCET-12-0919; The Fundamental Research Funds for the Central

Abstract

Recently, increasing numbers of users have adopted microblogging services as their main information source. However, most of them find themselves drowning in the millions of posts produced by other users every day. To cope with this, identifying a set of the most influential people is paramount. Moreover, finding a set of related influential users to expand the coverage of one particular topic is required in real world scenarios. Most of the existing algorithms in this area focus on topology-related methods such as PageRank. These methods mine link structures to find the expected influential rank of users. However, because they ignore the interaction data, these methods turn out to be less effective in social networks. In reality, a variety of topics exist within the information diffusing through the network. Because they have different interests, users play different roles in the diffusion of information related to different topics. As a result, distinguishing influential leaders according to different topics is also worthy of research. In this paper, we propose a multi-topic influence diffusion model (MTID) based on traces acquired from historic information. We decompose the influential scores of users into two parts: the direct influence determined by information propagation along the link structure and indirect influence that extends beyond the restrictions of direct follower relationships. To model the network from a multi-topical viewpoint, we introduce topic pools, each of which represents a particular topic information source. Then, we extract the topic distributions from the traces of tweets, determining the influence propagation probability and content generation probability. In the network, we adopt multiple ground nodes representing topic pools to connect every user through bidirectional links. Based on this multi-topical view of the network, we further introduce the topic-dependent rank (TD-Rank) algorithm to identify the multi-topic influential users. Our algorithm not only effectively overcomes the shortages of PageRank but also effectively produces a measure of topic-related rank. Extensive experiments on a Weibo dataset show that our model is both effective and robust.

Introduction

Online social networks provide a medium through which millions of users interact with each other; their members diffuse information and exhibit influence [1]. Influence analysis has



Universities under grant No. K5051203020, K5051303018, JB150313, JB150317, and BDY081422; Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase); Natural Science Foundation of Shaanxi Province, under grant No. 2010JM8027; The Creative Project of the Science and Technology State of xi'an under grant No. CXY1441(1); and The State Key Laboratory of Geoinformation Engineering under grant No. SKLGIE2014-M-4-4.

Competing Interests: The authors have declared that no competing interests exist.

received wide research attention. Among the investigated problems, finding influential individuals is an important topic for many applications such as online advertising, recommender systems and information diffusion. Consider online advertising as an example. When a new movie or product is released, the producer wants the item to be discussed frequently on social networks. The producer selects a small number of initial users from the network to "retweet" it (an action in the twitter-like social network that enables users to disseminate a certain item). These retweets are expected to launch gain a large amount of audience attention. The problem lies in determining who should be selected as the initial users to gain the most influence within the social network. Outside of advertising, when some emergency news needs to be announced about a particular subject, it is essential to select the most influential users as the seed users who will spread that news effectively. Both of these problems are related to social influence ranking, a problem that has attracted many studies.

Twitter-like social networks employ a network model called "following", in which each user is allowed to follow anyone without requiring permission. Based on these established social relations, a user will be alerted whenever a user they are following posts tweet updates [2]. Hence, this structure is the major means of influence propagation in these networks. Moreover, algorithms akin to PageRank [3] [4] have been used to find influential users on social networks whose topology is similar to the web. In PageRank, a random surfer is assumed to browse along links between web pages. However, most tweets are open-access for everyone in a Twitter-like social network, which means that users are able to retweet whatever they are interested in without the permission of the original poster. This open access breaks the restrictions of fixed social network structures. As a result, information freely propagates between users who may not necessarily have direct links in the social network. Most of the studies on social influence [5] [6] have targeted the actual retweet network instead of the "following" network. But topology is not the only means of information diffusion in social networks as it is in web networks. This paper decomposes the influential effects of a particular user into two parts: direct influence, which is triggered by a retweet action, and indirect influence, which is caused by the actions of others, without necessarily involving a following relationship.

In total, users generate an enormous number of tweets everyday on social media. Topic-mixed tweets propagate through the same network, and these various tweets contain the topic-related information. Faced with this flood of information, information seekers [7] are eager to follow only the most valuable users for a given topic. Actually, because sharing the same network structure, it is hard to distinguish different topics from the same information propagation. Discovering topic-dependent influential leaders enables user find the topic-related posters. On the other hand, Social network sites show new users a list of topics from which they can choose the ones they are most interested in following. Then, the most influential user for this topic will be recommended to the new user. Therefore, this "cold start" problem [8] could be well solved by recommending topic-related influential users. Importantly, influential users for one topic may fail to have the same influence for other topics, which means users play different roles in different topics [9]. Therefore, influence is topic-dependent, making it necessary to discover the topic-dependent influential leaders.

In this paper, we propose a novel influential model called the multi-topic influence diffusion (MTID) model to discover topic-related influential leaders. Specifically, the influence of these users in our model consists of both direct and indirect influence. The direct influence follows the information propagation trace along the links. Meanwhile, users can retweet tweets posted by people they are not following, giving the original posters further indirect influence. Notice that both types of influence are related to different topics. Based on MTID, we further propose a topic-dependent rank algorithm, namely TD-Rank. Different from setting one ground node to make the network strongly connected, as in LeaderRank [10] [11], we treat ground nodes as



the different topic pools of original tweets to construct a topical view of the whole network. Connecting all the nodes with each ground node hence represents a one-topic view of the network. This view is strongly connected and, thus, has the same properties as LeaderRank. Moreover, the transition probability in our model is extracted from traces of tweet action including both posting of the original tweet and retweeting others' tweets. This approach has been shown to be helpful in finding influential users [12]. We further experimentally demonstrate that our proposed ranking algorithm extracts nontrivial nodes as influential nodes in various topics on the large-scale Weibo network.

Related work

Generally, discovering influential leaders is related to two research topics. The first is measuring the maximization of influence. Influence maximization aims to find a set of seed users that influence a large number of other users. This problem was first studied by Domingos and Richardson from an algorithmic perspective [13]. Then, Kempe et al. [1] formulated the problem as a combinatorial discrete optimization problem; they proved that this problem was NP-hard and proposed a greedy algorithm to deal with it. Based on this basic discussion, many recent studies have also focused on topic-dependent issues. Chen et al. [14] studied topic-aware influence maximization; they proposed an algorithm to find k seeds from social network such that the topic-aware influence was maximized. Bakshy et al. [15] investigated the diffusion cascades generated by 1.6M Twitter users using a follower graph. They found that the number of followers was an important indicator of an influential user, further evidence that structure is important. Cha et al. [2] used a large amount of data collected from Twitter and presented an indepth comparison of three measures of influence: indegree, retweet and mentions. They made observations that most influential users can hold significant influence over a variety of topics, which also revealed the need to distinguish topic-dependent influence.

The second topic is the detection of authority nodes. One method for doing this is to create a rank list. Akin to the structure of the web, PageRank [3] [4] and the HITS algorithm [16] are also borrowed to investigate the influence-ranking problem. As a variant of PageRank, Lu et al. [10] proposed the LeaderRank algorithm to identify influential nodes by placing the ground node. Li et al. [11] further extended the algorithm by assigning degree-dependent weights to links associated with the ground node. All these methods determine influence based on a graph structure. By introducing the information propagation model [17], Zhu et al. [18] proposed a novel information diffusion model and integrated a Markov Chain into the independent cascade model. Based on this proposed model, they further proposed the rank algorithm SpreadRank to find influential users. To combine both greedy and heuristic algorithms, Cheng et al. [19] proposed IMRank, which found a self-consistent ranking by considering ranking-based marginal influence spread according to current ranking. TwitterRank [20] first measured the influence by taking both the topical similarity between users and the link structure into account. However, TwitterRank fails to distinguish the different types of topic-related influential leaders by assuming that tweets are retweeted according to a certain similarity. Some methods have tried to solve this problem with learning-based models. Su et al. [21] discussed the diversified expert-finding problem in academic social networks and proposed a new objective function to diversify the ranking list for a particular topic. Wang et al. [5] first introduced multi-task learning to predict individual influence based on the traces of information propagation.

Our work is different from all the studies described above in that we propose a novel influence diffusion model, to which we further add a novel topic-dependent rank algorithm. We introduce several ground nodes to decompose the total influence into direct influence and



indirect influence. The adoption of ground nodes allows the construction of multiple-topic views of the whole network. Moreover, this is helpful in finding topic-related influence. Meanwhile, it also overcomes the shortages of PageRank. In addition, we adopt a data-based approach to define the transition probability, which makes the model more accurate.

Materials and Methods

Multi-topic influence diffusion model

First, a directed social network $G = \{V, E\}$ is formed consisting of users and their following relationships, where V is the set of users and E represents the edge set of their followings. We denote $e = \langle i, j \rangle$ in E as a directed edge starting from i to j. Meanwhile users have traces of tweets $D = \{d_1, d_2, \ldots, d_N\}$, where d_i is trace of user i, and N is the number of users. We separate these tweets into two types: original tweets, those that are created originally by that user, and retweets, which are re-posted tweets created by others.

Based on the collected tweets, we conduct topic distillation, which aims to automatically identify the topics that users are interested in. For this purpose, the Latent Dirichlet Allocation (**LDA**) model [22] is applied. As a result, we denote P_i^o as the topic distribution of original tweets posted by user i, while P_i^r is user i's retweet topic distribution.

Using the above measurements, we describe the multi-topic influence diffusion model (MTID) in greater detail. First, suppose user j retweets one tweet from user i. Then, user i's influence is expanded by user j through user j's influence. We then set I(i) as the influence of user i. As stated above, we decompose the total influence of particular user as follows:

$$I(i) = I_d(i) + I_r(i), \quad i = 1, 2, ..., N$$
 (1)

where N is the number of users, $I_d(i)$ is the direct influence gained from followers, and $I_r(i)$ is the indirect influence caused by retweets from users who are not followers of user i. Moreover, taking the topic factor into consideration, we define the influence of topic t as $I^t(i)$. Thus, Eq. (1) can be rewritten as:

$$I^t(i) = I_d^t(i) + I_r^t(i) \tag{2}$$

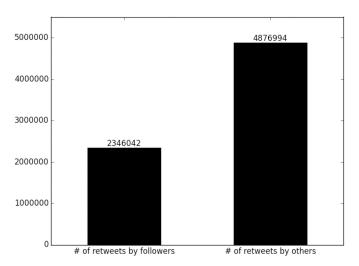
Before proceeding to deal with the definition of $I_d^t(i)$ and $I_r^t(i)$, we make two observations about Twitter-like social networks:

Observation 1: A user's posted tweets may be accessed by anyone in the whole social network due to the policies of online social network sites.

Observation 2: A follower's retweet action is the primary way by which users enlarge their scope of influence, which means that their influence diffuses as the tweet propagates along the network. More influential users will gain more retweets.

To verify these two observations, we calculated the statistics on our crawled data in Result Section, as illustrated in Fig 1.

In Fig 1(a), we count how many retweets are retweeted from followers and from other users who are not direct followers. Obviously, users can not only retweet tweets from people they are following but also those that originate from people they are not following. This statistic can validate *Observation 1*, and also clarifies the distinction between direct influence and indirect influence. We illustrate the distribution of retweet numbers per user in Fig 1(b). To show the distribution clearly, we take the log value of the number of users. As shown, most tweets are retweeted by fewer than 100 users. Only a few users have a larger retweet range, indicating that only the most influential leaders can gain large retweet numbers. This can be the consequence of *Observation 2*.



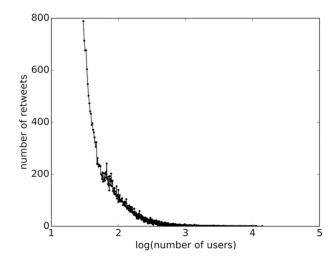


Fig 1. The statistics of Weibo data. (a) the retweet source distribution, (b) the number of retweets distributed over log(number of users).

doi:10.1371/journal.pone.0158855.g001

To explain *Observation 1*, we introduce *topic pools*. Topic pools are public information sources from which any user is able to find and retweet original tweets. This can be interpreted as a function of "*discovery*" (http://d.weibo.com/) in Twitter-like social network sites. The discovery function aims to enable users to find topic-related tweets despite the lack of existing following relationships.

Assume user i has a topic distribution P_i^o on his original tweets. The component of topic t is $p_{(i,t)}^o$. When user i posts an original tweet, that tweet contributes to the topic t pool with an expected value of $p_{(i,t)}^oI^t(i)$. As a result, the influence of each topic pool $I^t(g)$ consists of the combination of topic contributions from all the users:

$$I^{t}(g) = \sum_{i=1}^{N} p_{(i,t)}^{o} I^{t}(i)$$
(3)

According to Eq.(3), a user with higher influence will contribute more to the topic pool, indicating that more credits will be obtained if one tweet is retweeted by this user. Every topic pool is equally likely to distribute its influence among the whole network. Users are expected to gain $\frac{1}{N}I^t(g)$ influence from the topic pool. Therefore, the influence contributed by each user will be distributed among whole network via topic pools. Thus, we define the indirect influence of user i as:

$$I_r^t(i) = \frac{1}{N} I^t(g) \tag{4}$$

According to Eq (4), user i will gain influence indirectly through others without gaining any new follower relationships. Specifically, the influence of user i can expand rapidly when users with higher influence perform retweets.

Direct influence is gained from i's followers immediately. Suppose $F_i = \{f_1, \ldots, f_m\}$ denotes the follower list of user i, and m is the number of followers. Then, the direct influence can be defined as follows:

$$I_d^t(i) = \sum_{j \in F_i} p_{(i,j,t)}^r I^t(j)$$
 (5)



where $p_{(i,j,t)}^r$ is the retweet topic distribution between user i and user j on topic t. The Eq (5) explains the *Observation 2*. When user j retweets user i's tweets, the influence transits from j to i. Conversely, if j does not retweet any of i's tweets, the $p_{(i,j,t)}^r$ will be 0, meaning that user j only reads i's tweets but never retweets them.

Our **MTID** model is able to explain both the observations well from a data viewpoint. Note that our proposed model is different from the random surfer model in other algorithms; instead, our model is in accord with the available Tweet data and is not susceptible to the effects of manually determined parameters, which has further effects in the ranking algorithm based on the **MTID** model.

Topic Dependent Rank algorithm

Inspired by LeaderRank, we model topic pools as ground nodes inserted into the network as illustrated in $\underline{\text{Fig }2}$.

In the Fig 2, the ground nodes will establish bidirectional relations with each node. The network thus becomes strongly connected and consists of N + T nodes and $M + T^*N$ links, where

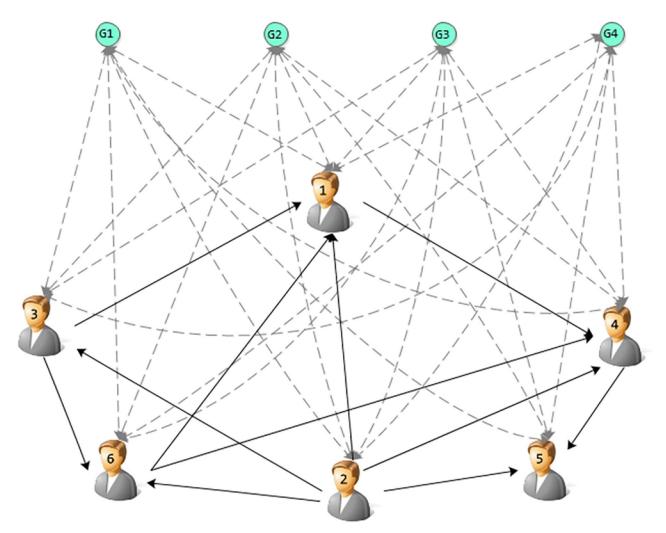


Fig 2. An illustration of the ground nodes. There are four ground nodes in this network representing four topics.

doi:10.1371/journal.pone.0158855.g002



N is the number of nodes, i.e users, M is the number of edges, and T is the number of topics. In addition, the weight is set as $p_{(i,t)}^o$ for the inbound direction from user i to ground node g representing topic t, while the outbound direction is $\frac{1}{N}I^t(g)$.

Therefore, we show the update rules for users' scores defined as $s_i = \{s_i^1, ..., s_i^T\}$ and the ground node scores on topic t as s_{σ}^t , respectively, as follows:

$$s_g^t = \sum_{i=1}^{N} p_{(i,t)}^o s_i^t \tag{6}$$

where s_i^t is the score of user i on the topic t. Moreover, we have

$$s_i^t = \sum_{j \in F_i} p_{(i,j,t)}^r s_j + \frac{1}{N} s_g^t \tag{7}$$

Notice that methods for calculating influential score by Eqs (6) and (7) are in accordance with the MTID model. In addition, we define the final score of a user to be the topic-dependent leadership score vector **S** including scores on all topics, namely,

$$\mathbf{S_{i}} = \mathbf{s_{i}}(Iter_{c}) + \frac{\mathbf{s_{g}}(Iter_{c})}{N}$$
 (8)

where $\mathbf{s_g}(Iter_c)$ is the vector expression of ground node scores on all of the topics at steady state, and $\mathbf{s_i}(Iter_c)$ is the vector expression of user *i*. The TD-Rank algorithm is summarized in Algorithm 1.

Algorithm 1 TD-Rank algorithm

```
Input: Network G = \{V, E\}, tweets associated with every user D = \{D_1, \ldots, D_N\},
  topic number T_i; the error threshold \epsilon to stop the iteration; the maximum
  iteration times Iter_{max}
Output: TD-Rank score of list TDS = [S_1, ..., S_N]
  process D with LDA according to the topic number T
  for t = 1 TO T do
    s_{\sigma}^{t}=0
  end for
  connect ground nodes to users with bidirectional edges
  set the weights on edges
  for i = 1 \text{ TO } N \text{ do}
    TDS[i] = S_i = [1/N, ..., 1/N]
  end for
  while err > \epsilon or k < Iter_{max} do
    Temp = TDS_k
    Update every ground node score according to Eq (6)
    Update every user score according to Eq (7)
    find the max error: err = Temp - TDS_k
  end while
```

In other algorithms akin to Pagerank, the final ranking equation can be defined as:

$$PR(i) = \alpha + (1 - \alpha) \sum_{\langle i,j \rangle \in E} p_{j,i} PR(j)$$
(9)

where PR(i) is the rank value of node i, α is the decay factor (i.e. the return probability), and $p_{j,i}$ is the transition probability matrix defined by the specific algorithm.



There are several drawbacks in applying this algorithm to social networks. First, the return probability is essential. Convergence is guaranteed only on strongly connected networks. In addition, the probability on every edge is identical for all users—irrespective of each user's tweet history. In comparison, our proposed TD-Rank based on the MTID effectively overcomes these shortages. Due to the adoption of ground nodes, TD-Rank extends the advantage of LeaderRank on every topic view. Moreover, we further adopt a data-driven approach to divide the transition probability into original and retweet probability. Finally, we reconstruct the network into a strongly connected network using the ground nodes and adopt a data-based approach to deal with the transition probability between users and ground nodes, aiming to discover an actual influence score.

Results

Datasets and experiment settings

To validate the effectiveness of the TD-Rank algorithm, we test it on crawled data from Weibo, the largest twitter-like social network in China. We start by randomly choosing several active seed users to avoid "Zombie users"—those who have registered but have not posted any tweets. Specifically, we include active users who retweeted more than 20 tweets between May 24th 2013 and May 24th, 2014. With these users, we crawl a network with 211,000 users, 1,612,289 following relationships and 47,002,906 total tweets. The detaileded statistics for this dataset are are listed in Table 1.

The only parameters that must be set are the LDA parameters, which reflect the number of ground nodes selected. The LDA is tuned by three parameters: the Dirichlet hyper-parameters α , β and topic number T. In this paper, these parameters are set as T=20, $\alpha=50/T+1$, and $\beta=0.1+1$ in Spark [23]. Obviously, choosing different values for these parameters has implications for the model results. However, this is a basically a model selection problem, which is not the focus of this paper. Then, we conduct LDA on traces of tweets D for every user. Table 2 lists 5 topics obtained by LDA as an example and the top 5 associated translated words in each topic.

Table 1. Dataset Statistics.

Users	Retweets	Original tweets	Relations
211,000	7,223,036	39,779,870	1,612,289

doi:10.1371/journal.pone.0158855.t001

Table 2. Examples of topics and associated words extracted by LDA.

Topic #	1	2	3	4	5
Associated words	company	user	children	designs	match
	management	technology	parents	photo works	sports
	marketing	intelligence	education	photography	the world cup
	brand	APPs	teacher	creativity	seasons
	market	Android	cultivation	style	NBA

doi:10.1371/journal.pone.0158855.t002



Experiment results

In this section, we make comparisons against related algorithms using the above dataset with Spark. The related algorithms studied include:

- PageRank, which measures the influence by taking the link structure of the network into account. The experiment setting in the comparison is set as follows: the return probability is 0.15 and the transition probability is $p_{ij} = 1/k_i^{out}$, representing the probability that i goes to j, while k_i^{out} denotes the outbound degree of i.
- **LeaderRank**, which introduced a ground node to connect all of the nodes, setting the transition probability as $p_{ii} = 1/k_i^{out}$.
- **TwitterRank**, which first studied topic-related ranking. In this comparison, we extract the same topics using LDA in TD-Rank.

The first experiment discusses the topic property of the rank result. We introduce the entropy on topic distribution. The entropy is calculated by:

$$E_{i} = -\sum_{t=1}^{T} p_{i,t}^{o} \log (p_{i,t}^{o})$$
(10)

Notice that we use the topic distribution on the original tweets here because original tweets are posted by a user to express his or her own interests. Then, we obtain the average entropy of users grouped in ranking list order as 1-10, 11-20, 21-30, 31-40 and 41-50. The comparison results are demonstrated in $\underline{\text{Fig 3}}$. For ease of visualization, we first compare PageRank, Leader-Rank and 20 topic-related TD-Rank results in $\underline{\text{Fig 3(a)}}$. For TwitterRank, we select top 10 users in every topic as an example and compare the results with TD-Rank in $\underline{\text{Fig 3(b)}}$.

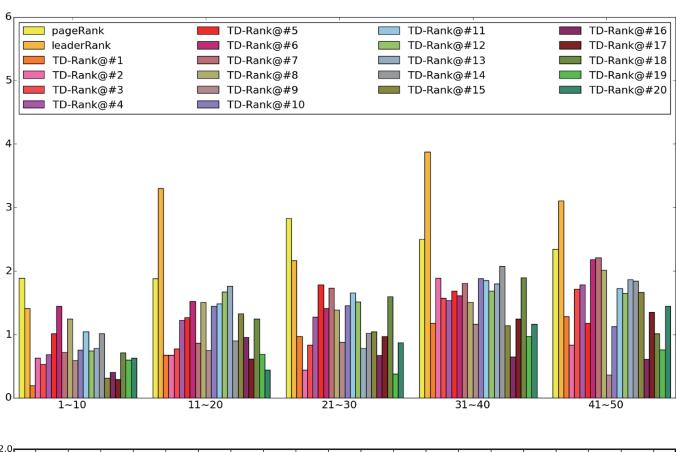
From the results, it is obvious that the users ranked by TD-Rank and TwitterRank have far less entropy compared with PageRank and LeaderRank, indicating that our proposed algorithm finds topic-related influential leaders similarly to TwitterRank. Moreover, in Fig 3(b) the ranking of users by TwitterRank has more entropy than the ranking by TD-Rank, indicating that users in our ranking list are more closely related to the same topic.

Another issue concerning the ranking results is the problem of robustness. Many spammers exist in social networks who attempt to gain reputation for advertising purposes [24]. To investigate this issue, we create the ν edges which link ν fake followers to every user and observe the positional changes in the ranking.

Specifically, we simulate the situation where a user creates v fake spammers and compare the positional changes in both ranking results. The whole process is described as follows. Suppose the user is i, we randomly select v users denoting as $\{u_1, u_2, \ldots, u_v\}$. Then following directed links are created to disturb the algorithm: $\{\langle u_1, i \rangle, \langle u_2, i \rangle, \ldots, \langle u_v, i \rangle\}$. The results are reported in Fig 4.

The horizontal axis of Fig 4 shows the original rank of a particular user, and the vertical axis is the manipulated rank after the addition of *v* spammers. As result, the diagonal line is the baseline, which indicates the rank is invariant. Vertical shift from the diagonal line corresponds to the change of ranking results. In addition, for ease of visualization, we select the ranking list on topic #1 as an example to show the comparative performance of TwitterRank and TD-Rank. As shown, TD-Rank is the most robust against spammers because the changes in ranking positions are much smaller than those of the other algorithms, and TwitterRank is more robust than PageRank and LeaderRank. Therefore, the results are mainly due to the distinguishing topics. In all, TD-Rank is a better algorithm for creating robust rankings in social network.





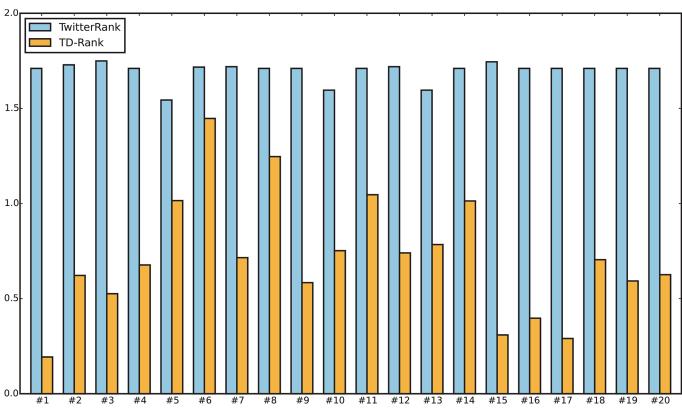




Fig 3. The topic entropy. (a) Comparison between PageRank, LeaderRank and TD-Rank; (b) Comparison between TD-Rank and TwitterRank on top 10 users.

doi:10.1371/journal.pone.0158855.g003

The next issue is influence maximization. We first consider two information-diffusing models in previous work [25].

Independent Cascade (IC) Model: This model begins with an initial set of active nodes. The process unfolds in discrete steps. When node v first becomes active in step t, it has a single chance to activate each currently inactive neighbor w based on parameter $p_{v,w}$. In our experiment, we set the parameter $p_{v,w}$ uniformly to 0.1.

Topic Independent Cascade (TIC) Model: This model uses the same process as IC, but its the parameter $p_{v,w}$ is related to topic. Specifically, the probability of diffusion is defined as:

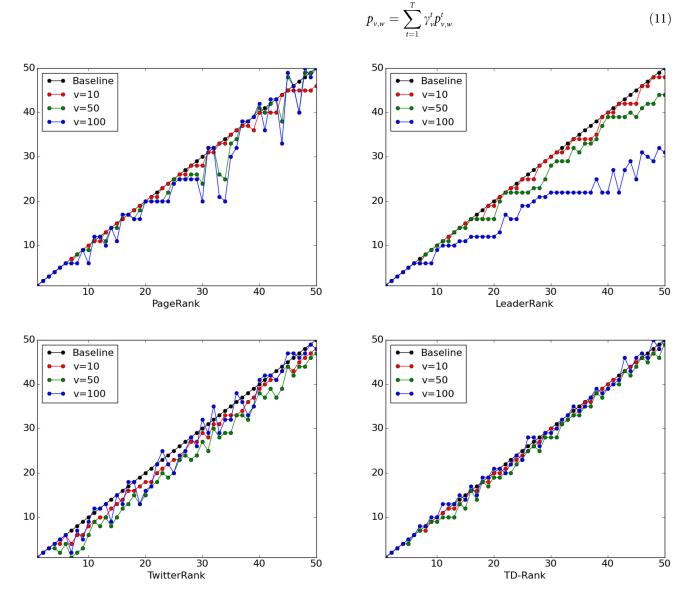


Fig 4. The spammer effect on ranking results. (a) PageRank, (b) LeaderRank (c) TwitterRank, (d) TD-Rank.

doi:10.1371/journal.pone.0158855.g004

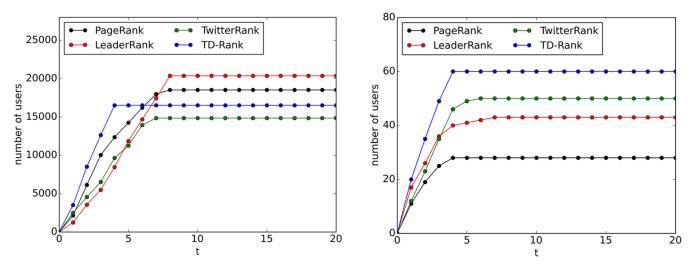


Fig 5. Comparison of influence maximization. (a) The results of IC Model. (b) The results of TIC Model.

doi:10.1371/journal.pone.0158855.g005

where γ_{ν}^{t} is the topic probability of user ν on topic t, and $p_{\nu,w}^{t}$ represents the influence strength exerted by user ν on w on topic t.

We conduct experiments on IC and TIC with the top 10 users in every ranking results consisting of the initial set of active nodes. In addition, the top 10 users in TD-Rank and Twitter-Rank are also selected from the ranking list of topic #1 as additional experiments. The results are shown in Fig 5.

Because it ignores the effects of topic distribution, the IC model is more suitable for a comprehensive ranking algorithm. However, it is worth noting that TD-Rank converges to a stable number faster than the others, which indicates that TD-Rank is more inclined to measure the influence of users on one particular topic with a smaller population. Compared to TIC, TD-Rank and TwitterRank calculate the influence on more users by considering topic effects. In all, this experiment shows the essential value of taking the topic factor into account.

We also extracted the number of retweets for several tweets posted by the top 50 users in the whole dataset. Then, we investigated the differences between this true retweet number and the number obtained by TIC. Assuming that the set of tweets posted by user i is denoted by d_i , we then evaluate the influence-predicting performances of all the algorithms using the mean squared error (MSE) as the evaluation metric:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{K} \left(\sum_{k \in d_i} (\hat{V}_i(k) - V_i(k))^2 \right) \right)$$
 (12)

where $\hat{V}_i(k)$ is the predicted number of retweets of a tweet k posted by user i, N is the number of users, and $V_i(k)$ is the real value.

The predictions' MSEs are presented in <u>Table 3</u>. Note that the settings for TwitterRank and TD-Rank are the same as in the other experiments.

Table 3. Comparison of the predictions' MSEs.

	PageRank	LeaderRank	TwitterRank	TD-Rank
MSE	29.95	27.21	18.03	16.38

doi:10.1371/journal.pone.0158855.t003



As shown in <u>Table 3</u>, PageRank and LeaderRank perform the worst, followed by Twitter-Rank. TD-Rank performs the best, outperforming the competition in terms of MSE, which indicates that the influential leaders discovered by our proposed algorithm are closer to the true situation.

Conclusion

In this paper, we focused on discovering multi-topic influential leaders in social network. We proposed a multi-topic influence diffusion (MTID) model, which decomposes the influence of a particular user into two parts: direct influence, which is influence related to that user's followers, and indirect influence, which is influence that is not restricted to direct followers. To cope with the definition of indirect influence, we introduced topic ground nodes that represent topic pools for establishing links between users. Moreover, to deal with the transition probability, we adopted a data-based approach that extracts the topic distribution from traces of tweets. Based on MTID, we further proposed a topic-dependent rank (TD-Rank) algorithm to identify the topic aware influential leaders. Finally, we conducted extensive experiments comparing the existing ranking algorithms using the Spark platform. The experimental results demonstrated that our proposed algorithm is more robust, more accurate and more sensitive to topic than previous algorithms.

Our plans for future work include dealing with the dynamic structure of a following social network by incorporating a time factor into our model. We will also consider other influence measures in the future.

Author Contributions

Conceived and designed the experiments: XT QGM. Performed the experiments: XT SSY. Analyzed the data: XT. Contributed reagents/materials/analysis tools: YNQ. Wrote the paper: XT. Collected the data: SSY.

References

- Kempe D, Kleinberg J, Tardos E. Maximizing the Spread of Influence Through a Social Network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2003. p. 137–146.
- Cha M, Haddadi H, Benevenuto F, Gummadi K. Measuring User Influence in Twitter: The Million Follower Fallacy. In: International AAAI Conference on Web and Social Media; 2010.
- Brin S, Page L. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report Stanford InfoLab. 1999;.
- 4. Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks. 2012; 56(18):3825–3833. doi: 10.1016/j.comnet.2012.10.007
- Wang Y, Xiang G, Chang SK. Sparse Multi-Task Learning for Detecting Influential Nodes in an Implicit Diffusion Network. In: AAAI Conference on Artificial Intelligence; 2013.
- Guo J, Zhang P, Zhou C, Cao Y, Guo L. Personalized Influence Maximization on Social Networks. In: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management. New York, NY, USA: ACM; 2013. p. 199–208.
- Guille A, Hacid H, Favre C, Zighed DA. Information Diffusion in Online Social Networks: A Survey. SIG-MOD Rec. 2013; 42(2):17–28. doi: 10.1145/2503792.2503797
- 8. Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems. Expert Systems with Applications. 2014; 41(4):2065–2073. doi: 10.1016/j.eswa.2013.09.005
- Chen C, Gao D, Li W, Hou Y. Inferring Topic-dependent Influence Roles of Twitter Users. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2014. p. 1203–1206.
- Lü L, Zhang YC, Yeung CH, Zhou T. Leaders in Social Networks, the Delicious Case. PLoS ONE. 2011 06; 6(6):1–9.



- 11. Li Q, Zhou T, Lü L, Chen D. Identifying influential spreaders by weighted leaderrank. Physica A: Statistical Mechanics and its Applications. 2014; 404:47–55. doi: 10.1016/j.physa.2014.02.041
- Goyal A, Bonchi F, Lakshmanan LVS. A Data-based Approach to Social Influence Maximization. Proceedings of the VLDB Endowment. 2011; 5(1):73–84. doi: 10.14778/2047485.2047492
- Richardson M, Domingos P. Mining Knowledge-sharing Sites for Viral Marketing. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2002. p. 61–70.
- Chen S, Fan J, Li G, Feng J, Tan KI, Tang J. Online Topic-aware Influence Maximization. Proc VLDB Endow. 2015; 8(6):666–677. doi: 10.14778/2735703.2735706
- 15. Bakshy E, Hofman JM, Mason WA, Watts DJ. Everyone's an Influencer: Quantifying Influence on Twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11. New York, NY, USA: ACM; 2011. p. 65–74.
- Kleinberg JM. Authoritative Sources in a Hyperlinked Environment. Journal of ACM. 1999; 46(5):604–632. doi: 10.1145/324133.324140
- Java A, Kolari P, Finin T, Oates T. Modeling the spread of influence on the blogosphere. In: Proceedings of the 15th international world wide web conference; 2006. p. 22–26.
- Zhu T, Wang B, Wu B, Zhu C. Maximizing the spread of influence ranking in social networks. Information Sciences. 2014; 278:535–544. doi: 10.1016/j.ins.2014.03.070
- Cheng S, Shen H, Huang J, Chen W, Cheng X. IMRank: Influence Maximization via Finding Self-consistent Ranking. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2014. p. 475–484.
- Weng J, Lim EP, Jiang J, He Q. TwitterRank: Finding Topic-sensitive Influential Twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York, NY, USA: ACM: 2010. p. 261–270.
- Su H, Tang J, Hong W. In: Learning to Diversify Expert Finding with Subtopics. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 330–341.
- 22. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003; 3:993–1022.
- Shanahan JG, Dai L. Large Scale Distributed Data Science Using Apache Spark. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2015. p. 2323–2324.
- 24. Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on twitter. In: Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS); 2010. p. 12–21.
- Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowledge and information systems. 2013; 37(3):555–584. doi: 10.1007/s10115-013-0646-6