Information Propagation in Microblog Networks

Chenyi Zhang[*†], Jianling Sun[*] and Ke Wang[†]
[*]College of Computer Science, Zhejiang University, China
[†]School of Computing Science, Simon Fraser University, Canada
chenyiz@sfu.ca, sunjl@zju.edu.cn, wangk@cs.sfu.ca

***Abstract*—*Information propagation* in a microblog network aims to identify a set of seed users for propagating a target message to as many interested users as possible. This problem differs from the traditional influence maximization in two major ways: it has a content-rich target message for propagation and it treats each link in the network as communication on certain topics and emphasizes the topic relevance of such communication in propagating the target message. In realistic situations, however, the topics associated with a link are not explicitly expressed but are hidden in the microblogs previously exchanged through the link. In this paper, we present a topic-aware solution to information propagation in a microblog network. We first model the latent topic structure of the network using observed microblog messages published in the network. We then present two methods for estimating the propagation probability based on the topic relevance between a link and the target message. Once the propagation probability is estimated, we adopt the standard greedy algorithm for influence maximization to find seed users. This approach is topic-aware in that the target message finds its way of propagation according to its topic relevance to the latent topic structure in the network. Experiments conducted on real Twitter datasets suggest that the proposed methods are able to select right seed users.**

I. Introduction

With the rapid growth of social network services and applications such as Facebook, Twitter and Weibo, research on social networks and social media is becoming a hot area. One example is social advertising [1], which utilizes user's relationships, interests and published data to target social advertisement to potential users. Microblogs, also called microposts, allow users to exchange small elements of content such as short sentences, individual images, or video links. Microbloggers post about topics ranging from the simple, such as "what I'm doing now," to the thematic, such as "sports cars".

A. Information Propagation

In this paper, we consider the problem of leveraging the abundant microblogs maintained by microblogging services to deliver some target information to microbloggers, or simply users. This problem, termed *information propagation*, can be stated as follows: given a microblog network with previously exchanged microblogs among users, we want to identify k seed users to propagate a target text message to as many users as possible in the network. The target message can be any text message such as a tweet, a web page, or an advertisement. Two related but different problems studied in the literature are *influence maximization* and *social contagion*. Influence maximization [2] aims to select some specified number of seed users that could influence the most number of users in a social network. Social contagion [3] refers to the phenomenon of information diffusion, such as diffusion of political opinions and adoption of new technologies [3]. Both problems leverage the link structure of a social network, but not messages exchanged among users, to influence more users. For example, traditional models of social contagion assume the probability that an individual is affected grows monotonically with the size of his or her neighborhood, and the recent study in [3] suggests that this probability grows with the number of connected components in the individual's neighborhood, not the size of the neighborhood.

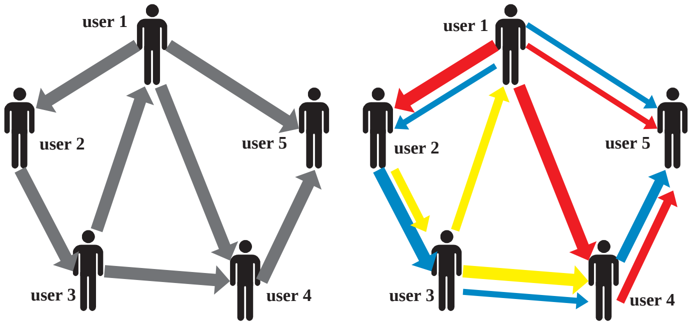


Fig. 1: Social network with simple links (left) and latent topics (right).

Information propagation differs from these existing problems in two major ways: it has a *content-rich target message* for propagation, and it treats each link in a microblog network as *communication on certain topics* and emphasizes the topic relevance of such communication when propagating the target message. The basic assumption in information propagation is that a target message is more likely to be forwarded or retweeted if it is interesting to both the sender and the recipient, and an interested user is more likely to react to a message (e.g., buying the advertised product). This assumption is consistent with previous studies that social influences are associated with certain topics [4] and marked tags or labels are useful for social interest discovery [5].

To illustrate the differences from influence maximization, Figure 1 shows a social network with link structure (left) and the network with links representing communication on certain topics (right), where each color represents a topic and the width of a link represents the intensity of the topic. Suppose that we want to propagate a target message on the topic corresponding to the yellow color, user 3 is more likely to be the best seed user to start the propagation because the message could reach two other users, namely user 1 and user 4. However, if this problem is treated as the traditional influence maximization, user 1 will be selected as the seed user because of its maximum out-degree, despite the fact that user 1 will not forward the message due to the lack of out-going communication on this topic. In this example, information propagation depends on not only the link structure of the network, but also the nature of a link in terms of the topics of the information exchanged.

To our knowledge, propagation of content-rich messages in a microblog network in a topic-aware manner has not been considered previously. The challenge is that the topics for messages and links are not explicitly expressed in a real life microblog network where only exchanged messages are observed. Manually labeling the topics for all messages and links, even for a training set, is unrealistic because expensive user involvement is required. The key to information propagation is to extract the hidden topics from the observable published messages in a microblog network and leverage them for identification of seed users.

B. Contributions

• We define the information propagation problem: *given a target message m and a positive number k, we want to identify k seed users in the microblog network for propagating m, with the goal of reaching as many users as possible*. Our assumption is that published microblogs *implicitly* convey the topics of communication represented by a link and that propagation of the target message depends on the topic relevance of such communication to the target message.

• We adopt the standard topic modeling technique, Latent Dirichlet Allocation (LDA) [6], to a microblog network to unveil the latent topics associated with social links. The outcome is a manifest of the topic distribution for each link, which serves as an explanation of the nature of a link in information flow. Our key insight is *not* applying LDA to the messages for each link individually, but to the whole collection of the messages for all links. We will explain the reasons for this approach.

• We propose two methods to estimate the propagation probability of a link based on the topic relevance between the target message and the link. Once propagation probability is estimated, we adopt the generic greedy algorithm for identifying seed users for the target message.

• We evaluate the performance of the proposed methods by comparing various probability estimation. Our study suggests that the proposed topic-aware propagation method selects more relevant seed users than traditional influence maximization methods.

In the rest of the paper, we review related work in Section II, present the topic modeling for microblog networks in Section III, present the propagation probability estimation and seed user selection in Section IV, and evaluate the proposed methods in Section V. Finally, we conclude the paper.

II. Related Works

A. Social Networks

One of the most robust findings in social networks is homophily [7] (i.e., “love of the same”), the tendency of individuals to associate and bond with similar others. Based on homophily, users tend to share interesting messages from their friends and spread from one person to another in the style of a biological epidemic. In addition to the link structure like in all social networks, a microblog network has its own characteristics, i.e., exchanges of abundant but short messages and interpersonal activities such as mentions and retweets. Such messages and activities convey certain important information about the users involved and play an important role in analyzing information diffusion in microblog networks. For example, [8] discussed information diffusion on Twitter via users’ ongoing social interactions, and [9] considered the content of retweet messages to improve topic mining for microblogs. To our knowledge, however, exploitation of published microblog messages for improving the propagation of a target message has not been formally studied.

B. Topic Modeling

Probabilistic topic models such as LDA were introduced by [6]. [10] presented the Author-Recipient-Topic (ART) model to learn the distribution specific to author-recipient pairs. [11] proposed a supervised learning approach to categorize links and quantify influence of web pages. Neither work considered information propagation. The supervised learning approach requires a training data set that is a link-labeled and link-weighted graph. Our work does not require such training data because it works directly on the microblog messages published by users.

Topic modeling has been used to predict social influences between users. [4] developed topical affinity propagation to model the topic-level social influence based on information of nodes, which was extended to heterogeneous networks in [12]. These methods assumed a *given* topic distribution for each node and found all topic level influence networks $G_z(V_z, E_z)$ for every topic z, where V_z is a subset of nodes that are related to topic z and E_z is the set of pair-wise weighted influence relations over V_z. These works did not consider propagation of information, which is the focus of our work. They assumed that the topic distribution is given for each user, whereas we assume that the topic distribution is hidden in the messages exchanged between users (thus, links). These works considered social influences for one topic at a time, whereas we treat each link as communication involving a topic distribution, instead of a single topic. The work in [13], [14] proposed a page rank based algorithm to find influential topics in twitter and citation network. Again, this work did not consider information propagation.

C. Influence Maximization

Influence maximization proposed in [2] aims to identify a set of seed users who could influence the most number of other users in a social network. Two popular influence propagation models are Independent Cascade Model and Linear Threshold Model. These models assume influence probability based on simple heuristics, such as uniform probability or probability proportional to the degree of a node. Moreover, this problem does not have a target message nor consider the topics for a link. Most previous works focused on improving the efficiency of greedy algorithms [15], [16], [17], [18], such as the CELF optimization based on the submodularity of incremental influences [15], [16].

Our work is closely related to the work on inferring the influence probability of a link. [19] used an “action log” to infer the influence probability of a link, where an action refers to a pre-determined activity such as joining a group. In the case that such actions are not explicitly captured in the social network, acquiring the action log requires the assistance from external information sources. Our work does not require such action logs. The works in [20], [21] used time decay to infer influence probability. Our work can be considered as a new way of estimating propagation probability by taking into account the topics of the microblog messages readily available in a microblogging service.

III. Topic Modeling For Microblog Networks

The first step of our method is to extract the latent topics on social links in a microblog network. We discuss first the extraction of microblog messages for a link and then topic modeling for such messages. The outcome is the topic distribution for each link and the word distribution for each topic. These distributions are used to estimate the propagation probability of a link for the target message in the next section.

A. Content based Social Links

There are two options for modeling the topics in a network: model the topics (i.e., interests) for each user, and model the topics for each link. Since we are interested in the topics for *relationships*, we adopt the second option. An example illustrates our choice. Consider two users A and B who are colleagues and have the same interests on three topics “work”, “travel” and “movie”. However, the two users have only exchanged the messages related to “work” (because they wish to limit their communication to work only). In this case, the user level topic modeling would suggest that these users will influence each other on the topics of “travel” and “movie” as well, which is a mistake because the two users did not communicate on the topics “travel” and “movie”. This example clearly shows that the link level topic modeling is a more natural choice for our purpose.

Microblog messages can be divided into three categories according to [22]: *broadcast messages*, *conversation messages*, and *retweet messages*. A broadcast message is published on the wall by some user A without a specific recipient. A conversation message is also published on the wall by a user A but a notification is sent to some user B to alert the publication. A message published by A is retweeted by a user B when B re-publishes the message on the wall, in which case the user A will get a notification that B has retweeted the message. All three types of messages can be viewed publicly, but only conversation and retweet messages involve an explicit information flow from a user A to a user B. Since a broadcast message has no explicit information flow between two users, it is hard to verify if any other user has an interest in a broadcast message. For these reasons, we shall use all three types of messages for topic modeling, but use only conversation and retweet messages to determine the topics for a link.

Conversation and retweet messages can be identified using special symbols within a message. For example, if the user A publishes a conversation message *“@B, Can you lend me a book on data mining”*, the user B will get a notification that A has published the message. This message flow can be observed on the social link $A \rightarrow B$. Similarly, from the retweet message published by the user A, *“Good job RT @B I have finished this experiment”*, the user B will get a notification that A has retweeted the message. This message flow can be observed on the social link $B \rightarrow A$. The structural symbol “@”, called *contactor factor*, indicates the contactor or recipient of a message, and the structural symbol “RT”, called *relation factor*, indicates the forward or quote relation [9]. Having clarified the above, we assume that each link e is associated with a set of conversation and retweet messages.

B. Topic Modeling for Links

Given a microblog network $G = (V, E)$, where V is the set of users and E is the set of social links, we want to determine the topic distribution for the communication represented by each link. Each link has a set of conversation and retweet messages, each being represented by a bag of words. Our approach is applying Latent Dirichlet allocation (LDA) [6] to the collection of microblog messages. Two issues must be resolved. The first is that each microblog message is very short (up to 140 characters), thus, sparse for topic mining. The second issue is that each social link may represent zero or more messages; dealing with each message individually leads to multiple topic distributions for each link, which is not only noisy due to the word sparsity of each message but also unrelated to each other due to the separate topic spaces. To address both issues, we model the set of messages associated with a link as one aggregated message by taking the union of these messages, and apply LDA to these messages plus all broadcast messages. Notice that there is only one topic modeling for the entire corpus, not one topic modeling per link.

Formally, let W be the size of the dictionary (i.e., the number of words) for messages, T be the number of latent topics, θ_m be the T-dimensional topic distribution for a message m, and φ_j be the W-dimensional word distribution for a topic j. The generative process of LDA is as follows:

1) choose $\varphi_j \sim Dir(\beta)$ where $j \in [1, ..., T]$
2) choose $\theta_m \sim Dir(\alpha)$ for each message m
3) for each word w_i that belongs to message m
    a) choose a topic $z_i \sim Mul(\theta_m)$
    b) choose a word $w_i \sim Mul(\varphi_{z_i})$

where β is the parameter of the Dirichlet prior on the per-topic word distribution and α is the parameter of the Dirichlet prior on the per-message topic distributions.

The Gibbs sampling [23] is widely used to infer the latent variables φ and θ (see [23] for the details of Gibbs sampling). Let $n_{j,w}$ be the number of times that the word w is assigned to the topic j in the sampling, $n_{m,j}$ be the number of times that the topic j is assigned to the message m in the sampling, $n_{m,\cdot}$ be the number of times that any topic is assigned to the message m in the sampling, and $n_{j,\cdot}$ be the number of times that any word is assigned to the topic j in the sampling. The topic distribution θ_m for a message m and the word distribution φ_j for each topic j are computed as follows:

$$\theta_m(j) = \frac{n_{m,j} + \alpha}{n_{m,\cdot} + T\alpha}, \quad \varphi_j(w) = \frac{n_{j,w} + \beta}{n_{j,\cdot} + W\beta} \qquad (1)$$

We can derive the topic distribution for links and the target message from θ and φ as follows:

Topic distribution for links: For each link e with the aggregated message m, the topic distribution of e, denoted θ_e, is defined as θ_m. A high value of $\theta_e(j)$ for a topic j indicates the existence of a *tunnel* for the communication on the topic j through the link e.

Topic distribution for the target message: For the target message m, the topic distribution of m is computed as follows. For each word w_i occurring in m, we determine the most likely topic for w_i, i.e., the topic j such that $\varphi_j(w_i)$ is maximal, and consider this as one vote for the topic j. Let $v_{m,j}$ denote the total number of votes for the topic j and let $\lambda_{m,j} = v_{m,j}/\sum_{i=1}^{T} v_{m,i}$, $1 \leq j \leq T$. The topic distribution of m is defined by $\lambda_m = \{\lambda_{m,1}, ..., \lambda_{m,T}\}$.

IV. Topic Aware Information Propagation

To achieve the goal of information propagation, seed users should likely publish the target message and the recipients of the message should likely forward the message, and so on. The likelihood of publishing or forwarding a message depends on whether there is a communication tunnel on the topics of the target message between the sender and the recipient. We can divide this problem into two sub-problems. The first sub-problem extracts the topic structure for links, which was addressed in the previous section. The second sub-problem will identify k seed users for propagating the target message, given the topic structure for links. We present three algorithms for the second sub-problem.

A. Greedy Algorithm: First Cut Solution

A first cut solution is ignoring all published messages and the target message and selecting seed users solely based on the topological structure of the microblog network. This is exactly the traditional influence maximization and a general greedy algorithm exists. This algorithm takes the graph structure of the microblog network and the number k as input and returns k seed users as output. Algorithm 1 below, *GeneralGreedy*, is an implementation of this algorithm from [2], [17]. It uses two internal parameters, the propagation probability P for all links e and the Monte Carlo random process of propagation starting from a set of users S, $MC(S, P)$. $MC(S, P)$ returns the estimated number of users reached by those in S. At each iteration, the algorithm greedily selects the next seed user v such that $MC(S \cup \{v\}, P)$ is maximized.

Algorithm 1 GeneralGreedy(G, k)

```
uniformly set propagation probability $P_e$ for all social links
$P = \bigcup_{e=1}^{|E|}\{P_e\}$
initialize $S = \emptyset$
for $i = 1$ to $k$ do
  select $v = \arg\max_{u \in V \setminus S}(MC(S \cup \{u\}, P))$
  $S = S \cup \{v\}$
end for
return $S$
```

Not surprisingly, the GeneralGreedy algorithm does not perform well for information propagation because it uses the uniform propagation probability P_e for all links e, which ignores the topic relevance of a link to the target message. Next, we present two topic-aware algorithms that take into account this topic relevance to infer the propagation probability P_e. These algorithms differ in the way of quantifying the topic relevance of a link.

B. Filtered Tunnel Algorithm

Let m denote the target message, e denote a link, and P_e denote the propagation probability of m through the link e. Recall that θ_e denotes the topic distribution of e and λ_m denotes the topic distribution of m. To determine P_e for a link e, we need to determine what topics of e are relevant to m. One way is cutting off insignificant topics in the topic distribution θ_e by a threshold, but this is not robust because it is difficult to know the proper threshold, which could vary from links to links. Our first topic-aware algorithm deals with this issue by classifying the topics for e as *tunneled topics* and *blocked topics*. The former refers to the topics that have large probabilities in θ_e to allow information flow on such topics, whereas the latter refers to the topics with insufficient probabilities to allow information flow. The intrinsic motivation for this classification is the observation that the distribution θ_e usually consists of a small number of major topics that have much higher probabilities than other topics.

To identify these two groups of topics, we apply the 2-means clustering method to the T data points represented by θ_e, where the jth point represents the probability on the topic j. The result is one cluster for tunneled topics and one cluster for blocked topics, represented by the indicator vector I_e:

$$I_e(j) = \begin{cases} 1 & \text{if } j \text{ is a tunneled topic} \\ 0 & \text{if } j \text{ is a blocked topic} \end{cases} \tag{2}$$

Notice that this classification of topics is on a per-link basis and is independent of the target message.

For a given target message m, we define the propagation probability for a link e as follows:

$$P_e(m) = \sum_{j=1}^{T} \lambda_{m,j}\theta_e(j)I_e(j) \tag{3}$$

In words, $P_e(m)$ is the inner product of the topic distribution of m and the topic distribution of e, except that only the topics j with $I_e(j) = 1$ have effect.

The seed user selection based on the above propagation probability, called *filtered tunnel algorithm* and denoted *FilteredTunnel*, is given in Algorithm 2. It takes a microblog network G, a positive number k, and the target message m as the input, and returns a set of k seed users as the output. The algorithm is an adaptation of the GeneralGreedy algorithm but uses the propagation probability $P_e(m)$ defined in Equation (3).

C. Unfiltered Tunnel Algorithm

The filtered tunnel algorithm adopts the "all or nothing" strategy for each topic on a link in order to focus on major topics. Sometimes two users communicate on a broad range of topics where there is no clear cut between tunneled topics and blocked topics. In such cases, a target message covering many topics could still be exchanged by users. This situation

Algorithm 2 FilteredTunnel(G, k, m)

```
for each social link $e$ do
    compute $P_e(m)$ as in Equation (3)
end for
$P = \bigcup_{e=1}^{|E|}\{P_e(m)\}$
initialize $S = \emptyset$
for $i = 1$ to $k$ do
    select $v = \arg\max_{u \in V \setminus S}(MC(S \cup \{u\}, P))$
    $S = S \cup \{v\}$
end for
return $S$
```

calls for the second approach, *unfiltered tunnel algorithm*, where topics with small probability are considered too for topic relevance. We can model this approach conveniently by making all topics the tunneled topics, that is, $I_e(j) = 1$ for every topic j in Equation (3), so the propagation probability $P_e(m)$ in Equation (3) degenerates into the usual inner product of the topic distribution λ_m of the target message m and the topic distribution θ_e of the link e:

$$P_e(m) = \sum_{j=1}^{T} \lambda_{m,j}\theta_e(j) \quad (4)$$

There are two cases for having a large propagation probability $P_e(m)$: either λ_m and θ_e have high probability in a few common topics, or λ_m and θ_e have small probability in many common topics. The former corresponds to communication on focused topics and the latter corresponds to communication on diversified topics. With $P_e(m)$ being defined by Equation (4), the unfiltered tunnel algorithm remains the same as Algorithm 2.

V. Experimental Evaluation

The ideal way of evaluating the propagation of a target message is placing the message to the selected seed users in a live microblogging service and tracing the propagation of the message. Unfortunately, this kind of evaluation requires full control over the microblogging service, which is possible only for the owner of a microblogging service. Without such full control over a microblogging service, we resort to publicly available Twitter microblog datasets[1] to approximate this evaluation. This dataset has over 9 million microblogs covering domains such as news, music, entertainment, technology, and web. We performed the following preprocessing: removed all users who have no social links and their broadcast messages because such users do not contribute to information flow; for the remaining users, took a random sample of their messages because topic modeling does not need all the data and running topic modeling on the whole collection of data is too slow; removed stop words and URLs from all messages. The final dataset contains 323481 messages (10% broadcast, 66% conversation, and 24% retweet), 10892 Twitter users, and 63454 links corresponding to followee/follower relations.

A. Experimental Design

We performed two experiments. In the first experiment, we randomly picked 50 target messages from relatively long retweet paths and withheld them from topic modeling and seed user selection. We study the *hit_ratio* of the seed users who published the *exact* target message, defined as the fraction of seed users who have forwarded the given target message according to the data set. While hit_ratio does measure the users who propagated the given target message, it does not consider the possibility of propagating any other messages, even such messages are similar to the target message. This exact syntax based measure could be too stringent because often a message is propagated because of its content, not because of its exact syntax. For example, if a user forwards the message *"@B, Does anyone know Canucks' standing"*, likely the user will also forward the message *"@B, Is Canucks in first or second place"*, if this message is presented instead. But the syntax based measure does not consider this flexibility.

In the second experiment, we relax the exact syntax requirement and consider two messages to be equivalent (with respect to propagation) if they are similar in topics. For two messages m_1 and m_2 with the topic distributions $\lambda_{m_i} = \{\lambda_{m_i,1}, \cdots, \lambda_{m_i,T}\}$, $i = 1, 2$, the *topic equivalence* of m_1 and m_2 is defined as

$$sim(m_1, m_2) = \sum_{j=1}^{T} \lambda_{m_1,j}\lambda_{m_2,j} \quad (5)$$

m_1 and m_2 are *topic equivalent* if $sim(m_1, m_2) > \varepsilon$ for some specified threshold ε. We consider the following two metrics based on topic equivalence. The *publish_ratio* is the fraction of the messages published by the seed users that are topic equivalent to the target message and the *reach_num* is the number of users reached through such forwarding. A larger value in these metrics means that a topic equivalent message is more likely to be published and propagated by the selected seed users. We randomly picked up 100 target messages for this experiment.

We evaluated three algorithms for information maximization.

GeneralGreedy, denoted GG: This is the traditional greedy algorithm in Algorithm 1, which was shown to outperform distance based, degree based, and random selection method [17]. We set P_e to 0.01, 0.02, 0.05, 0.1 as in [17]. GG0.01, GG0.02, GG0.05, and GG0.1 denote GG with these parameters.

FilteredTunnel, denoted FT: This is Algorithm 2. This algorithm used the hyperparameters α and β for topic mining. We set $\alpha = 1$ and $\beta = 0.01$ as in [24], and set $T = 50$ (the number of topics).

UnfilteredTunnel, denoted UT: This is the unfiltered tunnel algorithm described in Section IV.C. Like in FT, we set $\alpha = 1$, $\beta = 0.01$, and $T = 50$.

The number of seed users k is set to 10 and 50 for all three algorithms. We adopted a CELF optimization package[2] for the Monte Carlo random process $MC(S, P)$ in all three algorithms. This optimization speeds up the runtime but does not alter the result. All codes were written in Matlab and Java. The experiments were run on a PC with 3.10 GHz Quad-Core CPU, 8G memory and Operating System of Ubuntu Linux 9.10.

[1] http://user.informatik.uni-goettingen.de/~txu/cuckoo/dataset.html

[2] http://www.cs.ubc.ca/~goyal/code-release.php

TABLE I: Hit_ratio of GG, FT, and UT(%)

	GG0.01	GG0.02	GG0.05	GG0.1	FT	UT
$k = 10$	0.6	0	0	0	**0.8**	0.2
$k = 50$	0.36	0.2	0.12	0	**0.6**	0.28

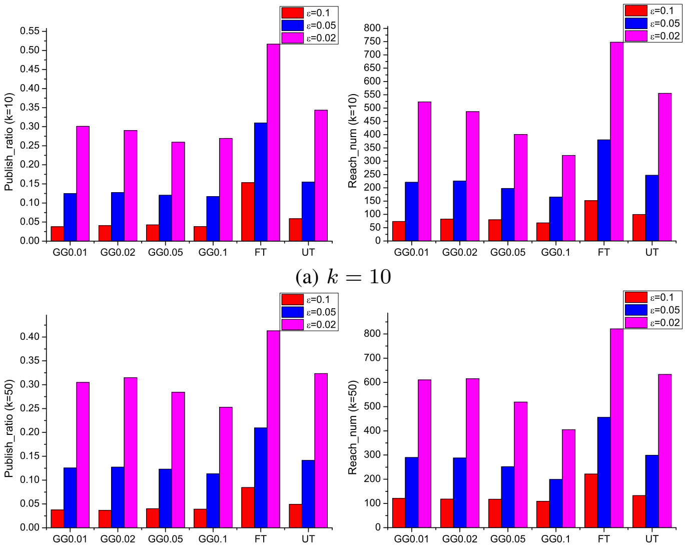
B. Evaluation based on Exact Messages

Table I shows the hit_ratio of GG, FT and UT (averaged over all target messages). Understandably, hit_ratio is rather low for all algorithms because only the users who published the exact target message are considered in this metric. Despite this, there is a notable difference among the three algorithms. GG is very sensitive to the setting of the propagation probability P_e. For the small propagation probability $P_e = 0.01$, GG0.01 tends to select central users in a dense community as seed users; such users usually have a higher degree, thus, are likely publishing the target message. This explains the higher hit_ratio. As the propagation probability increases, GG tends to select seed users who bridge different communities because of the increased reachability, but such users actually are less influential because the number of forwarding is very low. In contrast, UT and FT are able to select seed users based on the topics of the target message. Such users are likely to publish the target message. FT has a better performance (i.e., a higher hit_ratio) than UT because of its focus on major topics. See more discussions on this point below.

C. Evaluation based on Topic Equivalent Messages

Figure 2 shows publish_ratio and reach_num of GG, FT and UT. The three colors represent the three settings of the threshold ε for topic equivalence in Equation (5).

FT has significantly higher publish_ratio and reach_num than GG. This improvement comes from a better selection of seed users by considering the relevance of links to the target message. In particular, for a given target message, FT considers not only the link connection, but also whether similar messages were previously propagated through such links. As such, FT tends to select those users who are likely to publish the target message (i.e., a high publish_ratio) and consequently the target message can reach more users (i.e., a high reach_num).



(a) $k = 10$

(b) $k = 50$

Fig. 2: Publish_ratio and reach_num of GG, FT and UT

For a closer examination, Figure 3 shows the comparison of FT and GG0.01 at the individual target message level for the case of $k = 10$ seed users. For each target message, there is a point (x, y) where y represents the metric for FT and x represents the metric for GG0.01. A point above the diagonal line $y = x$ means that FT outperforms GG0.01 by having a higher publish_ratio and a higher reach_num. For nearly all target messages considered, FT outperforms GG0.01 through a higher value in both metrics. This suggests that FT selects more influential seed users than GG0.01. Another study, which is not shown here, showed that FT outperforms UT in these metrics. One reason is that UT keeps many minor topics that are insufficient to trigger publishing or forwarding of the target message. This study suggests that the focus on major topics in FT is an effective strategy.

We also studied the actual seed users selected. For discussion purpose, we consider the single topic target message m_1 containing the words for topic 50, and the mixed topic target message m_2 containing all of the words from topics 46 and 50. In general, the seed users selected by GG are central in dense parts of the network but may not be influential in the topics of the target message, in terms of the likelihood of published messages being forwarded by others, whereas the seed users selected by FT are more influential. The seed users selected by UT tend to be a mixture of those selected by GG and those selected by FT because UT not only considers topic relevance but also adds low propagation probability to each link.

Figure 4 shows the topic distribution of the messages published by seed users. For m_1 (on the left), which has the topic 50, the messages published by the seed users selected by FT have the highest probability for topic 50, followed by the messages of the seed users selected by UT, followed by the messages published by the seed users selected by GG0.01. For m_2 (on the right), which is on the topic 46 and the topic 50, the messages published by the seed users selected by FT have higher probabilities in both of these topics than those selected by UT and GG0.01.

To summarize, our study suggests that the topic-aware FT and UT perform better than the traditional topic-blind GG for information propagation: they tend to select right seed users, as demonstrated by higher probability of the target message being published (i.e., higher publish_ratio) and more users being reached (i.e., higher reach_num). The superiority of FT over UT suggests that taking all topics of messages into account does not necessarily yield better results; in fact, minor topics tend to mislead the selection of seed users. FT addresses this issue by focusing on major topics.

D. Runtime

Although our focus is on selecting more relevant seed users, the topic-aware selection also helps reduce the running time of the selection process. For FT and UT, topic modeling took about 2 minutes in our experiments. This step does not depend on the choice of the target message and was performed only once for all target messages. Table II shows the running

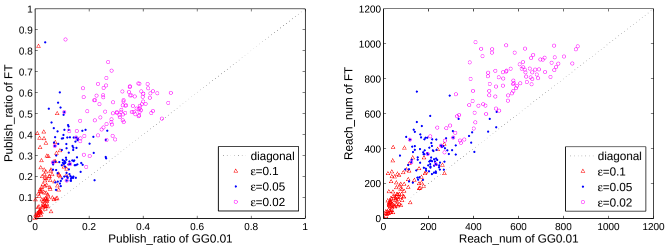


Fig. 3: Comparison of FT and GG0.01. $k = 10$

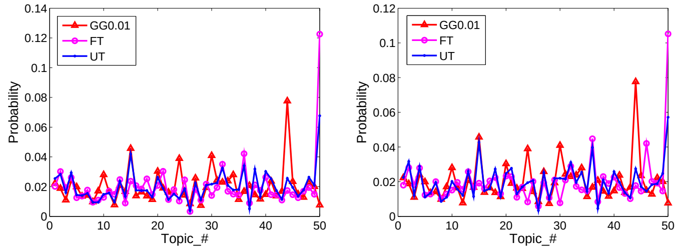


Fig. 4: Topic distribution of published messages of seed users for m_1 (left) and m_2 (right). $k = 10$

time for the selection of seed users. GG is highly sensitive to the choice of the propagation probability P_e because a larger probability means that GG will explore a larger part of the microblog network. For the topic-aware UT and FT, the running time is significantly reduced because propagation probability depends on the match between the topics of a link and the topics of the target message; consequently, only the links that are highly relevant to the target message are explored.

VI. Conclusion

This paper presented a study on propagating a target message to reach a maximal number of users in a microblog network. Existing solutions to influence maximization are not suitable for this problem because it does not factor the topic relevance of a link. Our contribution is a novel topic-aware estimation of the propagation probability of a link with respect to the target message. The novelty is that we do not assume that the topics of messages or links are given; rather, we assume that such topics are implicit in the microblogs published by microbloggers. We presented a method to extract such topics and use the extracted topics to infer the propagation probability for a target message. To our knowledge, this is the first work on estimating propagation probability in a topic-aware manner.

Acknowledgement

Jianling Sun's work is partially supported by Ministry of Industry and Information Technology of China (No. 2010ZX01042-002-003-001). Ke Wang's work is partially funded by a Discovery Grant from Natural Sciences and Engineering Research Council of Canada, and is partially done when he visited SA Center for Big Data Research hosted in Renmin University of China. This Center is partially funded by a Chinese National 111 Project Attracting International Talents in Data Engineering and Knowledge Engineering Research.

TABLE II: Average running time (min). $k = 50$

GG0.01	GG0.02	GG0.05	GG0.1	FT	UT
5.28	21.53	760.82	1242.57	2.35	33.57

References

[1] Y. Li and Y. Shiu, "A diffusion mechanism for social advertising over microblogs," *Decision Support Systems*, pp. 9–22, 2012.

[2] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *In KDD*, pp. 137–146, 2003.

[3] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[4] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," *In KDD*, pp. 807–816, 2009.

[5] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," *In WWW*, pp. 675–684, 2008.

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

[7] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[8] J. Yang and S. Counts, "Predicting the speed, scale,and range of information diffusion in twitter," *In ICWSM*, 2010.

[9] C. Zhang and J. Sun, "Large scale mircoblog mining using distributed mb-lda," *WWW(Companion Volume)*, pp. 1035–1042, 2012.

[10] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email," *J. Artif. Int. Res.*, vol. 30, no. 1, pp. 249–272, 2007.

[11] J. Tang, J. Zhang, J. X. Yu, Z. yang, K. Cai, R. Ma, L. Zhang, and Z. Su, "Topic distributions over links on web," *In ICDM*, pp. 1010–1015, 2009.

[12] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influecne in heterogeneous networks," *In CIKM*, pp. 199–208, 2010.

[13] J. Weng, E. Lim, J. Jiang, and Q. He, "Twitterrank : finding topic sensitive influential twitters," *In WSDM*, pp. 261–270, 2010.

[14] R. Nallapati, D. McFarland, and C. Manning, "Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents," *In AISTATS*, pp. 543–551, 2011.

[15] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection networks," *In kDD*, pp. 420–429, 2007.

[16] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Celf++: Optimizing the greedy algorithm for influence maximization in social networks," *In WWW(Companion Volume)*, pp. 47–48, 2011.

[17] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," *In KDD*, pp. 199–208, 2009.

[18] M. Mathioudakis, F. Bonchi, C. Castillo, and A. G. ans A. Ukkonen, "Sparsification of influence networks," *In KDD*, pp. 529–537, 2011.

[19] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *Proc. VLDB Endow*, vol. 5, no. 1, pp. 73–84, 2011.

[20] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *In KDD*, pp. 1019–1028, 2010.

[21] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," *In WSDM*, pp. 241–250, 2010.

[22] J. H. Kang, K. Lerman, and A. Plangprasophchok, "Analyzing microblogs with affinity propagation," *1st Workshop on Social Media Analytics(SOMA)*, pp. 67–70, 2010.

[23] T. L. Griffiths and M. Steyvers, "Finding secientific topics," *Proceedings of the National Academy of Sciences of the nited States of America*, vol. 101, pp. 5228–5235, 2004.

[24] T. Griffiths and M. Steyvers, "Probabilistic topic models," *Latent Semantic Analysis: A Road to Meaning. Hillsdale, NJ: Laurence Erbaum*, 2006.