CrossMark

Identifying and tracking topic-level influencers in the microblog streams

Sen Su[1] · **Yakun Wang**[1] · **Zhongbao Zhang**[1] ·
Cheng Chang[1] · **Muhammad Azam Zia**[1]

Abstract Topic-level social influence analysis has been playing an important role in the online social networks like microblogs. Previous works usually use the cumulative number of links, such as the number of followers, to measure users' topic-level influence in a static network. However, they ignore the dynamics of influence and the methods they proposed can not be applied to social streams. To address the limitations of prior works, we firstly propose a novel topic-level influence over time (TIT) model integrating the text, links and time to analyze the topic-level temporal influence of each user. We then design an influence decay based approach to measure users' topic-level influence from the learned temporal influence. In order to track the influencers in data streams, we combine TIT and the influence decay method into a united online model (named oTIT), which is applicable to dynamic scenario. Through extensive experiments, we demonstrate the superiority of our approach, compared with the baseline and the state-of-the-art method. Moreover, we discover influence exhibits significantly different variation patterns over different topics, which verifies our viewpoint and gives us a new angle to understand its dynamic nature.

Keywords Social influence · Graphical model · Online · Sina Weibo

✉ Sen Su
    susen@bupt.edu.cn

    Yakun Wang
    wangyakun@bupt.edu.cn

    Zhongbao Zhang
    zhongbaozb@bupt.edu.cn

    Cheng Chang
    changwow@bupt.edu.cn

    Muhammad Azam Zia
    zia@bupt.edu.cn

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

🖄 Springer

1 Introduction

1.1 Background and motivation

Microblogs, such as Twitter and Sina Weibo, have gained an explosive growth in popularity. Meanwhile, many applications emerge on these platforms, such as advertising and recommendation. For these applications, it is of great importance to have a better understanding of users' influence, since social influence captures the ways in which people affect others' opinions and behaviors (Zhang et al. 2013).

Users' influence usually varies in different topics (Cha et al. 2010). For example, Ming Yao is more authoritative on basketball rather than other aspects such as choosing a facial mask. Researchers have focused on topic-level influence analysis (Weng et al. 2010; Tang et al. 2009; Liu et al. 2010; Bi et al. 2014b; Tang et al. 2011). In general, people want to find current influencers rather than outdated ones.

Since, for the real applications like marketing, manufacturers usually select the current influencers as their product spokespersons. This makes sense because celebrities' influence changes over time (Cha et al. 2010; Wang et al. 2015; Goyal et al. 2010; Subbian et al. 2016) and it is challenging for influencers to maintain their status when many emerging local opinion leaders and evangelists enter the arena (Cha et al. 2010). Given the dynamics of influence, it is reasonably expected that users who attract audiences' attention at the present time are preferred for real applications compared with those who are attractive in the past. However, in the prior studies, it is common to utilize the cumulative number of social links (e.g., followship, reposts and mentions) to identify the topic-level influencers. As a result, they usually find faded influencers who were once popular but no longer attractive today, since they ignore the dynamics of influence and way they adopt is far from adequate. Accordingly, for measuring users' influence, it is critical to incorporate the variation trend of influence.

For example in Fig. 1, we can observe that user A attracts more and more followers, while user B presents the opposite case. Although they get the same number of followers at last, it is clear that user A is more attractive for the audience than user B over time. Given the dynamic nature of influence (Cha et al. 2010; Wang et al. 2015; Goyal et al. 2010; Subbian et al. 2016), to choose an influencer from them, there is no doubt that user A is preferred. A real example regarding two famous basketball players, Jianlian Yi and Jeremy Lin, in Sina Weibo is illustrated in Fig. 2. We can see that although Yi has more followers than Lin, the number of Yi's followers no longer increases, while Lin gets more and more followers along
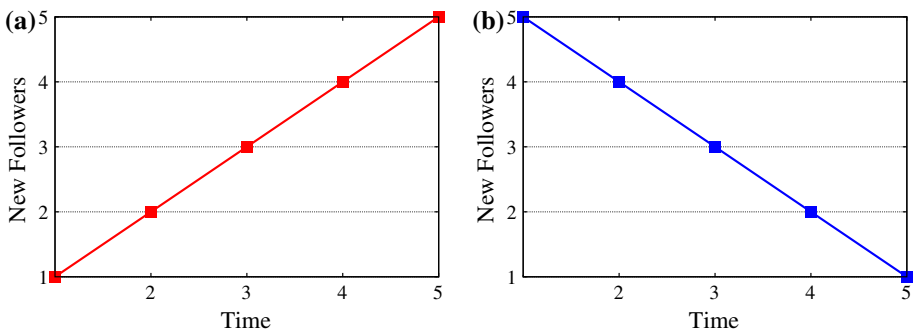
Fig. 1 The number of new followers **a** user A and **b** user B get over time
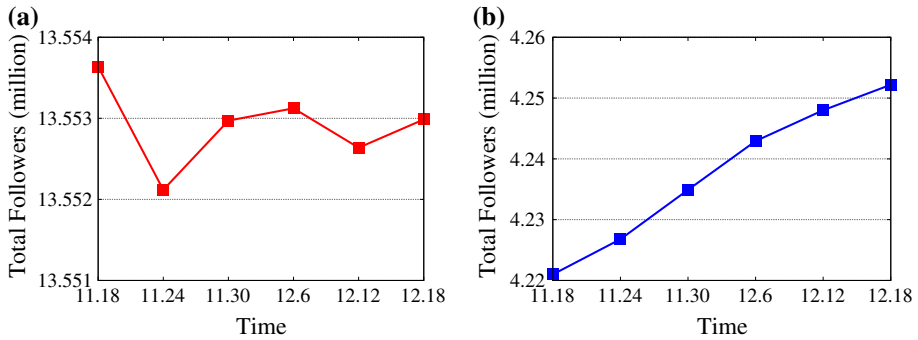
Fig. 2 The number of total followers over time of **a** Jianlian Yi and **b** Jeremy Lin in Sina Weibo (year 2015)

with time. This is reasonable since Yi was popular several years ago, especially when he played in the National Basketball Association (NBA). But after he left NBA and returned to Chinese Basketball Association (CBA), he lost attention gradually. On the contrary, Lin is now playing a critical role in the NBA from the start of "Linsanity". Accordingly, we can not simply assume Yi has more influence than Lin just because Yi owns more followers. However, if assuming both Yi and Lin are followed for basketball, all prior methods will select Yi as the key influencer rather than Lin, which leads to inaccurate models. This example conveys that the learned influence by the cumulative number of links is inadequate, since users' influence is dynamic and rises or falls over time (Cha et al. 2010). It is critical to integrate the temporal trends when measuring users' influence. Moreover, prior works usually analyze the influence in static networks (Weng et al. 2010; Tang et al. 2009; Liu et al. 2010; Bi et al. 2014b; Tang et al. 2011; Pal and Counts 2011; Bi et al. 2014a), which are inapplicable to real-world scenario.

1.2 Proposed solutions

In this paper, we study the problem of analyzing the topic-level temporal influence of users for identifying the key influencers on specific topics in the microblog sphere. Note that, the key influencers we intend to find are those popular and influential persons of the day rather than faded ones.

To address this problem, we firstly propose a novel probabilistic generative model, which we refer to as Topic-level Influence over Time, abbreviated as TIT. The basic idea of TIT comes from the answers of the following three questions: (1) How to detect topics on microblogs? (2) How to utilize the underlying network to measure users' topic-level influence when extracting topics? (3) How to incorporate the vibrant temporal factors to analyze influence? Firstly, we use the classical Latent Dirichlet Allocation (LDA) model (Blei et al. 2003) to detect topics of each user from their aggregated posts. Secondly, for modeling the topic-level influence, we let the links be generated by topics from the same topic space as words. As a result, the users co-followed by others with similar topical interests can be a good indicator of key influencers on these topics. Meanwhile, we use a Bernoulli distribution to capture the reasons why a user pays attention to another on microblogs. Besides the topic based reason, we group the topic independent links into another cluster as the topic unrelated influencers due to the fact that topic is not the only reason for users to follow others in microblogs (Bi et al. 2014b). Thirdly, for capturing the temporal influence, we associate links with time information. When a link is generated from TIT, the corresponding time is

generated simultaneously. Thereby, we can obtain a distribution of topics over links and time. With the learning result of this distribution by Gibbs sampling (Griffiths and Steyvers 2004), we can draw the changing influence of each user on specific topics over time like Fig. 2. Note that we address all these issues adequately in the TIT model. After that, we devise a method that takes the influence decay into consideration to compute the topic-level influence of each user.

Moreover, for adapting the model to the real-world data streams, we combine TIT with the influence decay method into a united online model: the oTIT model. oTIT is a sequential model over social streams each of which is a TIT model. Estimation results of parameters in the current stream are used as the prior for the next stream, and only newly arrived data needs to be sampled, which not only captures the temporal dynamics of influence but also greatly reduce the computation and storage requirement. Through extensive experiments on real-world dataset, we demonstrate the effectiveness and efficiency of our approach.

This paper extends our previous conference article (Wang et al. 2016) with the following improvements. (1) We give more information regarding the design of TIT model. (2) We introduce an online algorithm for TIT to handle large scale data sets. (3) More comprehensive experiments were conducted, and new findings are reported. (4) Derivations of collapsed Gibbs Sampling for TIT are provided.

1.3 Key contributions

- We are the first to propose to identify current topic-level influencers on Microblogs. We design a novel probabilistic generative model, called TIT, to jointly model the text, links and time in the microblog network for analyzing users' topic-level temporal influence. Meanwhile, we propose an influence decay method to measure the topic-level influence of each user based on the learned temporal influence, which takes both quantity and trend of influence into consideration.
- We combine TIT and influence decay into a united online model, named oTIT, to track the influencers in data streams, which not only captures the temporal dynamics of influence but also reduces the cost of time and memory.
- We conduct extensive experiments on a real-world dataset. Experimental results show that our approach significantly outperforms the baseline and the state-of-the-art algorithm by precisely identifying the topic-level key influencers in the microblog sphere.
- We are the first to discover that influence exhibits significantly different variation patterns on different topics. This interesting and valuable insight provides us a new angle of view to understand social influence and its dynamic nature.

1.4 Roadmap

The rest of the paper is organized as follows. Section 2 gives the problem statement. Section 3 presents the TIT model as well as an influence decay method, based on which we introduce the oTIT model. Our approach is evaluated in Sect. 4. In Sect. 5, we revisit the related work. At last, we conclude the paper and discuss about the future work in Sect. 6.

2 Problem statement

We give the notation used throughout this paper in Table 1.

Table 1 Notation

| Notation	Description
u	A microblog user
w	A word posted by a user
f	A user who u interacted with
t, s	A time slice and a stream
δ, δ'	Size of each time slice and stream
z	The topic assignment to a word
x	The topic assignment to a link
y	Binary indicator of whether u interacts with f based on u's topics
$\alpha, \beta, \gamma, \epsilon, \rho$	Parameters of the Dirichlet (Beta) priors on Multinomial (Bernoulli) distributions
θ	Per-user topic distribution
φ	Per-topic word distribution
μ	Per-user Bernoulli distribution over indicators
σ	Per-topic user distribution
π	Topic-unrelated distribution over users
U, K, W, T, S	Number of unique users, topics, words, time slices and streams
N_u, L_u	Number of words and links of user u
λ, λ'	Decay parameter for exponential decay and linear decay

Definition 1 (*Microblog Network*) A microblog network $G = (\mathcal{U}, \mathcal{W}, \mathcal{L}, \mathcal{T})$, where $\mathcal{U}$ is a set of U users, $\mathcal{W}$ is a set of words posted by $\mathcal{U}$, $\mathcal{L}$ is a set of directed links denoting the various types of user interaction such as followship, reposts and mentions, and $\mathcal{T}$ is a set of T time slices representing the generation time of $\mathcal{L}$. Let $u \in \mathcal{U}$ denote the tail and $f \in \mathcal{U}$ denote the head of a link. A directed link $(u, f) \in \mathcal{L}$ represents there exists communication from user u to f, e.g., u follows f.

Each user u in the network is associated with a set of posts, where each post contains a bag of words $w_u \in \mathcal{W}$ from a given vocabulary. By utilizing the topic model, e.g., Latent Dirichlet Allocation (LDA) (Blei et al. 2003), we can derive the latent topics.

Definition 2 (*Topic*) Based on the text content posted by users, a topic $k \in [1, K]$ is defined as a V-dimensional multinomial distribution φ_k over words. Each user u has a K-dimensional multinomial distribution θ_u over topics reflecting his/her topical interests.

On microblogs, there are many reasons for a user to pay attention to others, due to topical interests or just because someone is a famous person. Intuitively, the users co-followed by others with similar topical interests can be regarded as the influencers on these topics. Besides, users' influence is time-sensitive, which leads to the following definition.

Definition 3 (*Temporal Influence*) Each topic k is associated with a U-dimensional multinomial distribution σ_k over a set of 2-tuples $\{(f, t)\}$, where f is a user at the head of a directed link and t is the generation time of this link. The temporal influence of user f at time t can be defined as a function of $influence(f)@t = \sum_{k=1}^{K} g(\sigma_{f,t,k})$, where the function should
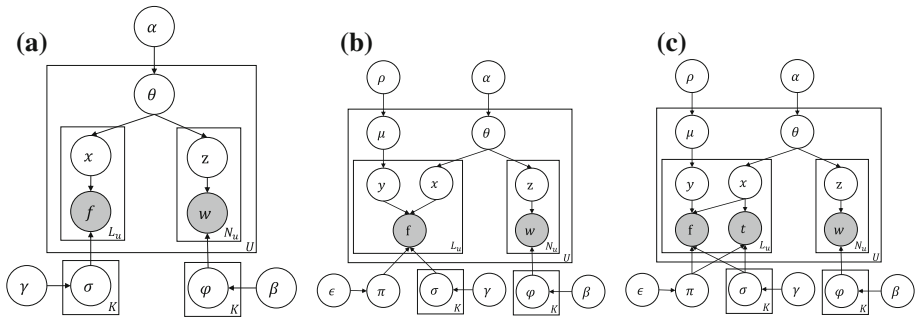
Fig. 3 Plate diagram of **a** Link-LDA, **b** FLDA and **c** TIT

capture two properties: 1) Quantity: the amount of attention f draws from other users. 2) Trend: how the amount of attention f draws varies over time.

The latent variable σ well captures the temporal influence, i.e., how users' influence changes over time. However, previous works ignore the variation trend and only use the cumulative number of links to measure users' influence. As a result, the influencers identified by them are those who accumulate the most attention on given topics, no matter whether they are still popular or not at the present time. Instead, the key topic-level influencers we aim to find are those who not just accumulate lots of attention from others, but especially have a growing trend of influence in recent time. Given the definition of temporal influence, we define the topic-level influencer identification and tracking problem.

Definition 4 (*Topic-Level Influencers Identification and Tracking Problem*) Given the microblog network $G = (\mathcal{U}, \mathcal{W}, \mathcal{L}, \mathcal{T})$ and the derived topic-level temporal influence σ, the influence of user f on topic k at time t can be defined as a function of $influence(f)@(k, t) = g(\sigma_{f,t,k})$. The key influencers $f \in \mathcal{U}$ we intend to find on topic k at time t should satisfy: $\forall f' \in \mathcal{U}, g(\sigma_{f,t,k}) \geq g(\sigma_{f',t,k})$.

3 Topic-level influence analysis

3.1 TIT model

Link-LDA (Erosheva et al. 2004) models the topic-level influence through a generative model that detects the topics and infers influence at the same time for citation and hyperlink network. FLDA (Bi et al. 2014b) intends to identify the topic-level influencers in microblog sphere based on Link-LDA by considering the fact that topical interests are not the only reason for users to follow others. Unfortunately, both of them ignore the temporal dynamics of influence and always find faded influencers, which limits the application of the proposed methods.

In order to model the temporal aspect of influence, we propose a topic-level influence over time (TIT) model jointly over text, links and time. It uncovers the latent topics and users' topic-level temporal influence in a unified way. For clarity, we show the plate notation of Link-LDA, FLDA and TIT in Fig. 3. Specifically, there are two components in this model: the user-word component in the right part and the user-(link, time) component in the left part of Fig. 3c.

The user-word component is to model u's words. We aggregate the words w posted by u into an integrated document from which we use an LDA-based model to discover the latent topics. As a result, each user has a Multinomial distribution θ over topics and each topic has a Multinomial distribution φ over words.

The user-(link, time) component is to model the u's links and the corresponding generation time in the microblog network. We discretize the time by dividing the entire time span of all links into T time slices. We consider the network as a document corpus and each user u is represented by a document where the links (i.e., the users f that u communicated with) and the corresponding time t pairs form the words in this document. Note that this component consists of two levels of mixtures: an upper-level Bernoulli mixture μ and two underneath-level multinomial mixture parts σ and π. μ is for deciding whether the link creation is based on u's topics or not. If topic based, we model the topic x (generated by θ) over (f, t) by a multinomial distribution σ. Otherwise, we use a global multinomial distribution π to model (f, t). Different from FLDA, TIT not only models the links in the network but also the corresponding time about when the links emerge. As a result, σ in TIT captures the temporal influence of users on specific topics. Benefiting from the learning results of σ, we can generate the influence trend line over time of each user like Fig. 2, and this can greatly help us to identify the key topic-level influencers on microblogs.

The generative process is summarized in Algorithm 1. Consider a user u who publishes a word $w_{u,m}$. He first selects a topic $z_{u,m}$ by his user-topic distribution θ_u and then selects the word by the topic-word distribution $\phi_{z_{u,m}}$. On the other hand, he creates a link to $f_{u,l}$ at time $t_{u,l}$. For generating $f_{u,l}$ and $t_{u,l}$, he firstly uses a Bernoulli distribution μ_u to generate a binary indicator parameter $y_{u,l}$ to decide whether $f_{u,l}$ is related to a topic. When $y_{u,l} = 1$, he creates a link to $f_{u,l}$ based on his topical interests. He uses the user-topic distribution θ_u to generate a topic $x_{u,l}$. Then a topic-(link, time) multinomial distribution $\sigma_{x_{u,l}}$ is selected to generate $f_{u,l}$ and $t_{u,l}$. When $y_{u,l} = 0$, $f_{u,l}$ and $t_{u,l}$ are generated by π for the topic-unrelated reason.

Algorithm 1: Generative process for TIT

Sample $\pi \sim$ Dirichlet (ϵ)
for *each topic $k \in [1, K]$* **do**
  Sample the distribution over words $\varphi_k \sim$ Dirichlet (β)
  Sample the distribution over links and time $\sigma_k \sim$ Dirichlet (γ)

for *each user $u \in [1, U]$* **do**
  Sample the distribution over topics $\theta_u \sim$ Dirichlet (α)
  for *the mth word of user u, where $m \in [1, N_u]$* **do**
    Sample the topic $z_{u,m} \sim$ Multinomial (θ_u)
    Sample the word $w_{u,m} \sim$ Multinomial $(\varphi_{z_{u,m}})$
  Sample the distribution over indicators $\mu_u \sim$ Beta (ρ)
  for *the lth link with corresponding time t of user u, where $l \in [1, L_u]$* **do**
    Sample the topics $x_{u,l} \sim$ Multinomial (θ_u)
    Sample the indicator $y_{u,l} \sim$ Bernouli (μ_u)
    if $y_{u,l} = 1$ **then**
      Sample the link and time $(f_{u,l}, t_{u,l}) \sim$ Multinomial $(\sigma_{x_{u,l}})$
    if $y_{u,l} = 0$ **then**
      Sample the link and time $(f_{u,l}, t_{u,l}) \sim$ Multinomial (π)

3.2 Parameter estimation

Since exact inference for TIT model is intractable, we therefore utilize collapsed Gibbs sampling (Griffiths and Steyvers 2004), a widely used Markov Chain Monte Carlo (MCMC) algo-

rithm, to obtain samples of the hidden variable assignment and to estimate the model parameters from these samples. Gibbs sampling iteratively samples latent variables (i.e., z, x, y in TIT) from a Markov chain, whose stationary distribution is the posterior. We provide the sampling equation below, and the detailed derivation of these equation is given in the "Appendix".

Let $z_{\neg j}$ denote the set of all hidden variables of topics except z_j and $n^{(\cdot)}_{\cdot,\neg j}$ denote the count that the element j is excluded from the corresponding topic or user. We use similar symbols for other variables. Firstly, we sample the topic assignments $z_{u,m}$ for $w_{u,m}$ with index $j = (u, m)$ given the observations and other assignments using a Gibbs sampling procedure in Eq. 1:

$$
\begin{aligned}
&p(z_j | z_{\neg j}, x, w, \alpha, \beta) \\
&\propto \frac{n^{(w)}_{k,\neg j} + \beta}{\sum_{w=1}^{W} n^{(w)}_{k,\neg j} + W\beta} \left(n^{(k)}_{u(w),\neg j} + n^{(k)}_{u(f)} + \alpha \right),
\end{aligned}
\tag{1}
$$

where $n^{(w)}_{k,\neg j}$ refers to the number of times that word w has been observed with topic k, $n^{(k)}_{u(w),\neg j}$ denotes the number of times that topic k has been observed with a word w of user u, and $n^{(k)}_{u(f)}$ denotes the number of times that topic k has been observed with a link f of user u.

Then, for a user $f_{u,l}$ at the head of a link and the corresponding time $t_{u,l}$ with index $i = (u, l)$, we jointly sample y_i and x_i from the conditional in Eqs. 2 and 3:

$$
\begin{aligned}
&p(x_i, y_i = 1 | f, t, x_{\neg i}, y_{\neg i}, z, \alpha, \gamma, \rho) \\
&\propto \frac{\sum_{t=1}^{T} n^{(f,t)}_{k,\neg i} + \gamma}{\sum_{f=1}^{U} \sum_{t=1}^{T} n^{(f,t)}_{k,\neg i} + U\gamma} \left(n^{(y=1)}_{u,\neg i} + \rho_1 \right) \left(n^{(k)}_{u(w)} + n^{(k)}_{u(f),\neg i} + \alpha \right)
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
&p(x_i, y_i = 0 | f, t, x_{\neg i}, y_{\neg i}, z, \alpha, \epsilon, \rho) \\
&\propto \frac{\sum_{t=1}^{T} n_{(f,t),\neg i} + \epsilon}{\sum_{f=1}^{U} \sum_{t=1}^{T} n_{(f,t),\neg i} + U\epsilon} \left(n^{(y=0)}_{u,\neg i} + \rho_0 \right) \left(n^{(k)}_{u(w)} + n^{(k)}_{u(f),\neg i} + \alpha \right),
\end{aligned}
\tag{3}
$$

where $n^{(f,*)}_{k,\neg i}$ denotes the number of times that user f occurs in topic k, $n_{(f,*),\neg i}$ denotes the number of times that user f occurs without any topic, $*$ represents an aggregation on time dimension, $n^{(y=1)}_{u,\neg i}$ and $n^{(y=0)}_{u,\neg i}$ denote the number of times the links created by u is related to topics or regardless of topics, respectively.

After a sufficient number of iterations, we can estimate the unknown parameters based on the samples by Eqs. 4–8.

$$
\vartheta_{u,k} = \frac{n^{(k)}_{u(w)} + n^{(k)}_{u(f)} + \alpha}{\sum_{k=1}^{K} \left(n^{(k)}_{u(w)} + n^{(k)}_{u(f)} \right) + K\alpha}
\tag{4}
$$

$$
\varphi_{k,v} = \frac{n^{(w)}_{k} + \beta}{\sum_{w=1}^{V} n^{(w)}_{k} + V\beta}
\tag{5}
$$

$$
\mu_{u,y} = \frac{n^{(y)}_{u} + \rho_y}{n^{(y=1)}_{u} + n^{(y=0)}_{u} + \rho_0 + \rho_1}
\tag{6}
$$

$$
\sigma_{f,k} = \frac{\sum_{t=1}^{T} n^{(f,t)}_{k} + \gamma}{\sum_{f=1}^{U} \sum_{t=1}^{T} n^{(f,t)}_{k} + U\gamma}
\tag{7}
$$

$$\pi_f = \frac{\sum_{t=1}^{T} n_{(f,t)} + \varepsilon}{\sum_{f=1}^{U} \sum_{t=1}^{T} n_{(f,t)} + U\varepsilon} \tag{8}$$

3.3 Measuring users' temporal influence

Given the topic-level influence trend lines over time derived from σ, users who get lots of attention from others and have a upward trend of influence can be easily found as the key influencers on the corresponding topics. However, in some cases like the example Yi and Lin in Fig. 2, we can not easily identify who exhibits more influence, since Yi has more followers than Lin, while Lin has a better growing trend of influence than Yi. Intuitively, the links generated long time ago have little contribution to users' influence. It means the more closer of the links generated in time, the more important they are to users' influence. Hence, we utilize the exponential decay function to model the influence decay. Specifically, σ is a distribution of topics over a set of 2-tuples $\{(f, t)\}$. That is, σ is a $U \times T \times K$ matrix in the procedure of sampling recording the number of times (f, t) has been assigned to topic k, denoted as $n_k^{(f,t)}$, plus prior parameter γ, i.e., $\sigma_{u,t,k} \propto n_k^{(f,t)} + \gamma$. Thus, we can use Eq. 9 to measure the topic-level temporal influence of user f on topic k till time T:

$$Influence(f)@(k, T) = \gamma + \sum_{t=1}^{T} n_k^{(f,t)} \times e^{-\frac{T-t}{\lambda}} \quad \lambda > 0. \tag{9}$$

Here, λ is a parameter controlling the decay rate of influence.

3.4 Online TIT model

We have introduced the TIT model that models the temporal aspect of topic-level influence. In addition, an exponential decay based method is proposed to compute users' influence on specific topics. In this section, we combine TIT and influence decay into a united online model, named oTIT, for tracking the topic-level influencers in data streams, which prior methods fail to do.

The oTIT model is inspired by the online LDA model proposed in AlSumait et al. (2008), which assumes the data arrives in a streaming fashion in ascending order of their publication date. The main idea of oTIT is to fit a TIT model over the data in a stream $s \in [1, S]$ and use the counts in current stream to adjust the hyperparameters for the next stream $s + 1$. For example, the count of word w in topic k, i.e., $(n_k^{(w)})^{(s)}$, resulted from running TIT at stream s, can be used to update the prior hyperparameter β for $s + 1$, and likewise for other hyperparameters. It means we use the historical assignments of latent variables as prior observations for next incoming stream. Here, we use δ' and δ to denote the size of each stream s and time slice t. In general, a stream contains at least one time slice, i.e., $\delta' \geqslant \delta$. t is the generation time of links, and δ can be a minute, an hour or a day. The setting of δ' depends on how fine or coarse the results are expected to be in specific applications. For example, Sina Weibo provides monthly updated ranking lists of influencers on specific topics. In such case, δ' can be predetermined as a month. Prior online and dynamic models (Blei and Lafferty 2006; AlSumait et al. 2008) treat the documents in each stream exchangeable, which faces the problem that long time window length may lead to losing the track of highly dynamic influence. In contrast, we consider the time sequence of links in each stream, which allows oTIT to be a more fine-grained model to capture the variation of influence. Besides, for new words and new links in new incoming streams, their prior counts are set to default values in oTIT. The updating formulas of hyperparameters are shown in Eqs. 10–14.

$$\alpha_{u,k}^{(s+1)} = \alpha_{u,k}^{(s)} + \left(n_{u(w)}^{(k)}\right)^{(s)} \times e^{-\frac{\delta'}{\lambda'}} \tag{10}$$

$$\beta_{k,v}^{(s+1)} = \beta_{k,v}^{(s)} + \left(n_k^{(w)}\right)^{(s)} \times e^{-\frac{\delta'}{\lambda'}} \tag{11}$$

$$\rho_{u,y}^{(s+1)} = \rho_{u,y}^{(s)} + \left(n_u^{(y)}\right)^{(s)} \times e^{-\frac{\delta'}{\lambda'}} \tag{12}$$

$$\gamma_{k,f}^{(s+1)} = \gamma_{k,f}^{(s)} + \sum_{t=(s-1)\times\delta'}^{s\times\delta'} \left(n_k^{(f,t)}\right)^{(s)} \times e^{-\frac{s\times\delta'-t}{\lambda}} \tag{13}$$

$$\varepsilon^{(s+1)} = \varepsilon^{(s)} + \sum_{t=(s-1)\times\delta'}^{s\times\delta'} (n_{f,t})^{(s)} \times e^{-\frac{s\times\delta'-t}{\lambda}}, \tag{14}$$

where $\lambda', \lambda > 0$, which are parameters controlling the decay of effect of historical learnt parameters. Here we use exponential decay to model the decay of historical influence. Next, we prove that we incorporate influence decay the same way as Eq. 9 into oTIT.

For Eq. 13:

$$
\begin{aligned}
\gamma_{k,f}^{(s+1)} &= \gamma_{k,f}^{(s)} + \sum_{t=(s-1)\times\delta'}^{s\times\delta'} \left(n_k^{(f,t)}\right)^{(s)} \times e^{-\frac{s\times\delta'-t}{\lambda}} \\
&= \gamma_{k,f}^{(1)} + \sum_{t=1}^{\delta'} \left(n_k^{(f,t)}\right)^{(1)} \times e^{-\frac{\delta'-t}{\lambda}} + \sum_{t=\delta'}^{2\delta'} \left(n_k^{(f,t)}\right)^{(2)} \times e^{-\frac{2\delta'-t}{\lambda}} \\
&\quad + \cdots + \sum_{t=(s-1)\times\delta'}^{s\times\delta'} \left(n_k^{(f,t)}\right)^{(s)} \times e^{-\frac{s\times\delta'-t}{\lambda}} \\
&= \gamma_{k,f}^{(1)} + \sum_{t=1}^{s\times\delta'} n_k^{(f,t)} \times e^{-\frac{s\times\delta'-t}{\lambda}}.
\end{aligned}
$$

Let $T = s \times \delta'$, apparently,

$$Influence(f)@(k, T) = \gamma_{k,f}^{(s+1)}.$$

The prominent advantages of oTIT in comparison to prior approaches (e.g., Link-LDA and FLDA) are in that, firstly, we utilize the temporal dynamics of influence, which plays a critical role in identifying current influencers. Secondly, we consider a dynamic scenario, i.e., data streams, instead of statics for tracking the influencers, which allows oTIT more efficient in terms of time and memory cost and more applicable to real-world scenario. The overall procedure of oTIT is shown in Algorithm 2.

3.5 Application

Besides marketing and recommendation, identifying current topic-level influencers in microblogs can benefit some other applications, e.g., feed ranking. Here we demonstrate how our model can be applied to it.

Feed ranking is to re-rank the items that users receive for satisfying their favor. For example in microblogs, we can select the top posts from the followees to show to users when they log

---

Algorithm 2: oTIT

Input: $K, \alpha^{(1)}, \beta^{(1)}, \rho^{(1)}, \gamma^{(1)}, \epsilon^{(1)}, \{w^{(s)}|1 \leqslant s \leqslant S\}, \{(f,t)^{(s)}|1 \leqslant t \leqslant T, 1 \leqslant s \leqslant S\}$
Output: $\{\theta^{(s)}, \phi^{(s)}, \mu^{(s)}, \sigma^{(s)}, \pi^{(s)}|1 \leqslant s \leqslant S\}$
Set default values for $\alpha^{(1)}, \beta^{(1)}, \rho^{(1)}, \gamma^{(1)}, \epsilon^{(1)}$
for *stream s = 1 to S* **do**
  Initialize $y^{(s)}, x^{(s)}$ and $z^{(s)}$ with random assignments
  $[z^{(s)}]$=GibbsSampling($w^{(s)}, \alpha^{(s)}, \beta^{(s)}$)
  $[y^{(s)}, x^{(s)}]$=GibbsSampling($f^{(s)}, t^{(s)}, \rho^{(s)}, \gamma^{(s)}, \epsilon^{(s)}$)
  Compute $\theta^{(s)}, \phi^{(s)}, \mu^{(s)}, \sigma^{(s)}, \pi^{(s)}$ by Equation $4 - 8$.
  Update $\alpha^{(s+1)}, \beta^{(s+1)}, \rho^{(s+1)}, \gamma^{(s+1)}, \epsilon^{(s+1)}$ by Equation $10 - 14$.

---

into the microblog system. For our concerned user u with topical interests distribution θ_u at time t, we generate a candidate set of interested followees for u as follows:

$$p(f|u, \theta, t) \propto \sum_{k=1}^{K} \theta_{u,k} \times influence(f)@(k, t), \tag{15}$$

which based on the interests of user u and the popularity degree of his followee f at time t. Then we sort all the candidate followees by the probability and show the recent ranked posts of them to u.

4 Experiment evaluation

4.1 Dataset

For comparing with the previous work, we crawl the followship network from Sina Weibo[1] which is one of the most popular microblog platforms. In Sina Weibo, users post text messages (up to 140 characters) to express their ideas and interests. Users interaction, e.g., follow relationships, imposes an underlying social network. Since Sina Weibo does not release the information about when a user follows another, we periodically crawl the follow list of all users in our seed set, monitor their changes and then label the new generated links with timestamps. We also crawl the recent 100 messages posted by users. This dataset is crawled between December 1st, 2015 and January 5th, 2016. We choose this dataset for the rich microblog text and the millions of links along with it. Finally, after appropriately preprocessing, there are 0.4M users, 207M words, 46M links with 7M time-tagged and 24 time slices with each nearly 1.5 days in our dataset. Naturally, $\delta = 1.5$ days and timestamp t ranges from 1 to 24, where 24 denotes the most recent time slice. We let each stream contain 4 time slices, i.e., $\delta' = 6$ days. Correspondingly, s ranges from 1 to 6. For old links without time information, we label them with random values from -400 to 0, and this part of data is considered as stream $s = 0$. Although the random assignments inevitably bring some noise, the experimental results still demonstrate the superiority of our approach.

4.2 Experiment setup

A significant step in such parametric method is choosing the proper hyperparameter values. We empirically set the number of topics $K = 100$ and the hyperparameters $\alpha = \frac{50}{K}, \beta =$

---

[1] http://weibo.com.

$\gamma = \epsilon = \tau = 0.01$, $\rho = 1$ (default values for oTIT). We set $\lambda' \to +\infty$ in oTIT, so that the words in different streams contribute equally. We set $\lambda = 11$ through minimizing held-out perplexity on a validation set. We run the Gibbs sampling for 500 iterations.

We evaluate our approaches by comparing them with Link-LDA (Erosheva et al. 2004) and Followship-LDA (FLDA) (Bi et al. 2014b). Recall that these two methods have been explained in detail in Sect. 3.1. In addition, for precision comparison, we implement a straightforward method that trains FLDA on each stream independently and then use a sum of topic-level influence on different streams as users' final influence score on corresponding topics. We refer to this model as FLDA-Stream. Another straightforward method is that we rank users based on the number of new followers they get in the past month after they are clustered into different topics. We refer to this method as New-Follower. We conduct the experiments on a server with 24 Intel(R) Xeon(R) CPU, 128 GB memory, 1.1T disk and CentOS release 6.4.

In the following section, we start with a qualitative analysis of our method through a case study and human judgement. Then we quantitatively evaluate the performance of our approach and the competitors in terms of precision and efficiency. Finally, we analyze the sensitivity of λ in terms of precision.

4.3 Qualitative analysis

4.3.1 Case study

Table 2 lists some of the resulting topics denoted by top keywords and corresponding top 5 ranked influencers by oTIT, FLDA and Link-LDA, respectively. As shown in this table, on the movie topic, oTIT not only identifies the most followed director Zhangke Jia and actor Kun Chen, but also a new famous director Hu Guan since he directs the movie "Mr. Six" which is highly praised and extremely popular in that period of time. As another example, on the music topic, oTIT identifies the most followed music media Netease Cloud Music and the current popular singer Dongye Song as well as the rock band Miserable Faith simultaneously. On the stock market topic, oTIT identifies current active and authoritative users who are well-versed at stock investment besides those most followed accounts like China Securities Regulatory Commission. Similar results are also gained on the topics sport and idol. Nevertheless, FLDA and Link-LDA only find the celebrities and organizations that have large number of followers on these topics, even they are not popular any more nowadays. It is clear that our approach produces significantly higher quality results than the competitors. In addition, it is worth mentioning that, on the sport topic, both FLDA and Link-LDA give Yi a higher rank than Lin. Although, oTIT also ranks Yi higher than Lin, the ranking of Lin in oTIT is much higher than that in FLDA and Link-LDA, which indicates oTIT tends to identify the increasingly popular users. Besides, some interesting phenomenons are gained from the experimental results. For example, users who are keen on horoscopes tend to follow the entertainers.
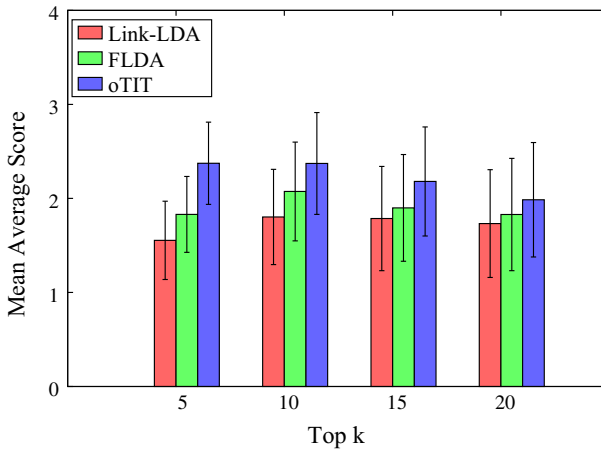
4.3.2 Human judgement

For further comparison, we resort to another method through human judgment to evaluate the performance of different approaches (Tang et al. 2009; Diao et al. 2012), because a good result should at least conform to peoples subjective judgement. We select top 20 influencers on each topic achieved by different approaches and mix them together into one list, respectively. Hence, we have 100 mixed lists corresponding to 100 topics with each no more than 60 non-anonymous influencers in it. Then we ask 10 graduate students to manually label the influencers by assigning a score (3: excellent, 2: good, 1: normal, 0: poor). We try to assign

Table 2 Top 5 influencers achieved by oTIT, FLDA and Link-LDA on 5 different topics

Keywords	Model	Top-5 influencers
Movie, director, billion, box office, release, cinema, starring, audience	oTIT	Zhangke Jia, Sina Movie,Hu Guan, Kun Chen, Xiaogang Feng
	FLDA	Zhangke Jia, Changwei Gu , Changtian Wang, Sina Movie, Kun Chen
	Link-LDA	Zhangke Jia, Changtian Wang, Sina Movie, Wei Zhao, Zhonglei Wang
Investment, stock market, economics, market, fund, capital, Shanghai composite index	oTIT	Shanshi Cao, China Securities Regulatory Commission, Daxiao Li, Rong Hua, Tianjin Guxia
	FLDA	Bin Dan, Daxiao Li, China Securities Regulatory Commission, CCTV Finance, Sina Finance
	Link-LDA	Bin Dan, Pi Shui, China Securities Regulatory Commission, Sina Finance, Daxiao Li
Music, band, music festival, song, rock, show, album, guitar	oTIT	Netease Cloud Music, Miserable Faith Band, Xiaosong Gao, Nanjing Li Zhi, Dongye Song
	FLDA	Xiaosong Gao, Zuoxiao Zuzhou, Xiami Music, Netease Cloud Music, Jun Zheng
	Link-LDA	Yadong Zhang , Xiaosong Gao , Xiami Music, Netease Cloud Music, Musician
Match, football, sport, club, final, fans, team, player	oTIT	Lin Gao, Guangzhou Evergrande, Jianxiang Huang, Jinyu Li, Yunlong Xu
	FLDA	Jianxiang Huang, Qiaosheng Han, Lin Gao, Lu Dong, Jinyu Li, Haidong Hao
	Link-LDA	Manchester United club, Jianxiang Huang, Gao lin, Barcelona club, Sina Sport
Fans, TV play, prince charming, vote, idol, celebrity, support, Ge Hu	oTIT	Ge Hu, Kai Wang, Hong Yuan, Yifeng Li, TFBOYS-Wang Yuan
	FLDA	Ge Hu, Hong Yuan, Yifeng Li, Yifei Liu, Tianyu Ma
	Link-LDA	Ge Hu, Hong Yuan, Nan Pai San Shu, Yifeng Li, Tianyu Ma

topics to students familiar with them and ask them to make a judgement based on their background knowledge as much as possible. They are allowed to consult the external resources to help their judgement. The criterion is to what extent the user is popular on corresponding topics, especially in recent time rather than early time. They can also use the total number of followers and the influence trend line of each user we provide for reference. They are asked to make judgements independently. In the end, we use the Mean Average Score assigned to each user as their final score, which is to eliminate the difference caused by subjective judgment as much as possible. Figure 4 shows the Mean Average Score of different methods across all topics over four settings of k with standard errors as error bars. It is clear that oTIT consistently gets higher scores than the others since oTIT successfully identifies the current influencers, while the competitors fail to do so.

Although these qualitative studies give us an intuitive appreciation that oTIT produces better results than the competitors, it still calls for a precision comparison.

Fig. 4 Mean Average Score of human judgement for different approaches across all topics over different settings of k. *Error bars* are standard errors

4.4 Quantitative analysis

4.4.1 Precision

For precision comparisons, there are no recognized topic-level rankings can be employed as the ground truth. Sina Weibo offers the rankings of popular users or organizations over 36 categories such as finance, sports and music, etc, and each category list contains 100 ranked users. These rankings imply some valuable information we need. Firstly, these popular users or organizations are some kind of the key influencers on the corresponding topics. Secondly, Sina Weibo states that these lists are updated by month. Intuitively, oTIT considering the temporal dynamics of influence should produce more precise results than the competitors. Although these rankings do not necessarily have 100% precision, they give us valuable information to facilitate relative comparisons across different approaches. Thus, we use these rankings as the ground truth to evaluate the performance of different methods in terms of (1) Hit Count at k ($HC@k$): the number of correctly detected top k users returned by a method among the ground truth; (2) Mean Average Precision at k ($MAP@k$): the proportion of correctly detected top k users among the ground truth across all categories.

Figure 5a, c, e show the results of all methods in terms of Hit Count over different settings of k. It is observed that our approaches significantly outperforms the competitors in terms of precision. In particular, FLDA-Stream performs worse since it loses significant temporal information of links and prior information in each running phase, which results in inaccurate model for influence analysis due to the information loss. The method of New-Follower is no better than FLDA-Stream since it only considers the new increasing followers and ignores the significant historical information. FLDA is superior to Link-LDA, since FLDA relaxes the assumption in Link-LDA that users follow others all because of topical interests. oTIT performs best among them, since oTIT not only correctly models the topic-level influence but also leverages its temporal dynamics. As a result, oTIT accurately identifies the current topic-level influencers. Due to the same decay schema used in both TIT and oTIT, they produce the similar results. Besides, we show the results in terms of mean average precision in Table 3.
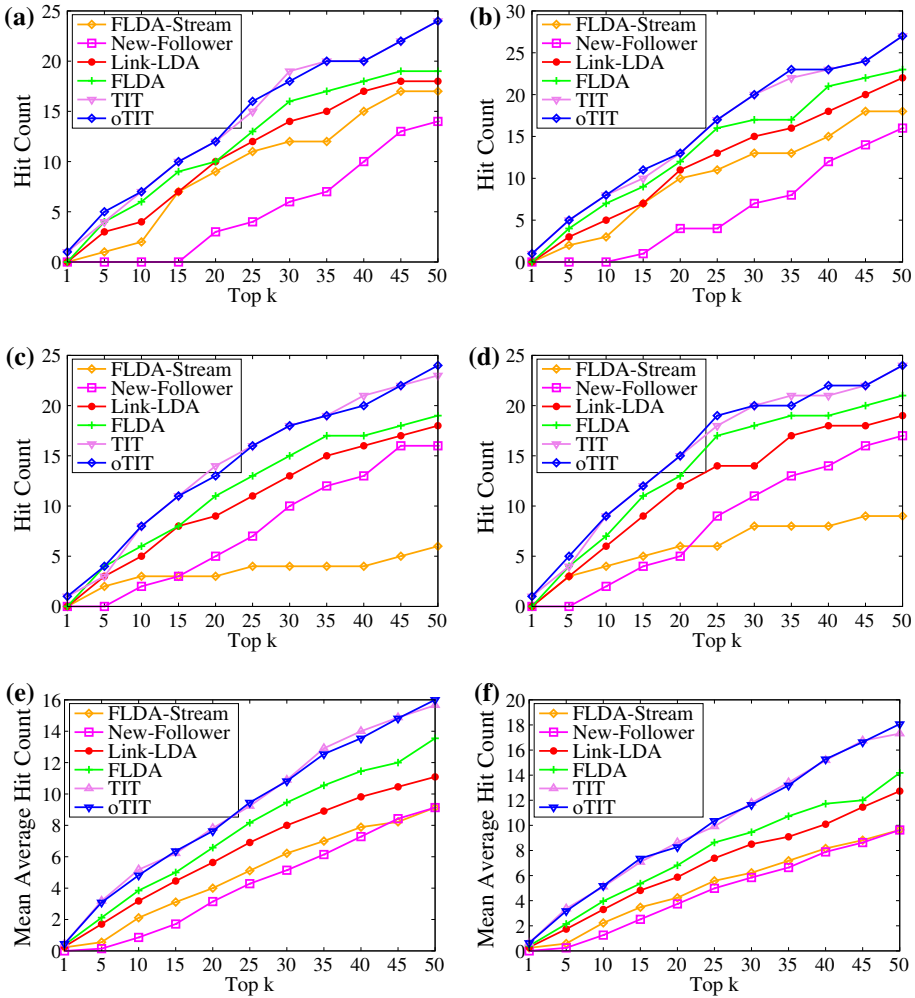
Fig. 5 Hit Count comparisons of all approaches over different settings of k. **a** Medical category, **b** medical category (large dataset), **c** movie category, **d** movie category (large dataset), **e** all categories and **f** all categories (large dataset)

Not surprisingly, oTIT consistently produces better precision results than the others, which verifies our analysis above.

More than that, we also conduct the precision comparison on a larger dataset with 1.1M users, 415M words and 98M links with 12M time-tagged. Although the this dataset spans the same period as prior one, it contains more users and follow relationships. For clarity, we show the details about these two datases in Table 5. Note that, dataset 1 is a part of dataset 2 and dataset 1 is used unless otherwise specified in our experiment. The comparison results in terms of Hit Count and MAP on this large dataset are reported in Fig. 5b, d, f and Table 4, respectively. We can observe that all approaches work better than on dataset 1. This may because some important influencers and network structures are included in this large dataset, which benefits the performance of different methods. Still, our approaches work the best among all methods.

Table 3 Mean average precision (*MAP*) comparisons of all approaches over different settings of *k*

Method	MAP@5 (%)	MAP@10 (%)	MAP@20 (%)	MAP@30 (%)	MAP@40 (%)	MAP@50 (%)
FLDA-stream	11.1	21.1	20.0	20.7	19.7	18.2
New-Follower	2.8	8.5	15.7	17.1	18.2	18.2
Link-LDA	34.1	31.8	28.1	26.6	24.5	22.1
FLDA	42.3	38.5	32.9	31.5	28.6	27.1
TIT	63.6	48.1	38.2	36.3	35.0	31.3
oTIT	61.8	51.8	39.1	36.1	33.9	32.1

Table 4 Mean average precision (*MAP*) comparisons of all approaches over different settings of *k* (large dataset)

Method	MAP@5 (%)	MAP@10 (%)	MAP@20 (%)	MAP@30 (%)	MAP@40 (%)	MAP@50 (%)
FLDA-stream	11.6	22.3	21.2	20.7	20.4	19.3
New-follower	4.8	12.6	18.7	19.5	19.7	19.2
Link-LDA	34.7	33.0	29.4	28.3	25.2	25.5
FLDA	42.9	39.6	34.1	31.5	29.3	28.4
TIT	65.2	61.8	41.4	39.3	37.9	34.6
oTIT	66.6	61.1	43.2	38.8	38.2	36.1

Table 5 Statistics of experimental dataset

Dataset	Users	Words	Links	Time-tagged links
dataset 1	0.4M	207M	46M	7M
dataset 2	1.1M	415M	98M	12M

Due to the similar results achieved by TIT and oTIT, we'd like to find out how much data is needed for oTIT to perform similarly to TIT. We run oTIT on the data from stream 1 to stream 6, streams 3 to stream 6 and stream 5 to stream 6, respectively, which are referred as oTIT-1, oTIT-2, and oTIT-3. We run TIT on the data from stream 1 to stream 6. Note that, stream 6 is the most recent data. Figure 6 shows the results of Mean Average Hit Count across all categories over different settings of *k*. It is observed that data loss damages the performance of oTIT.
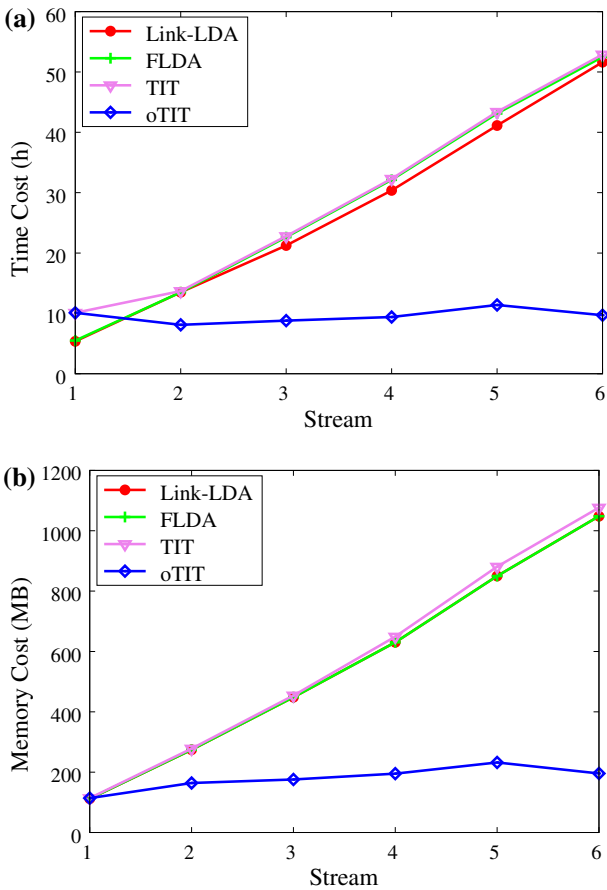
4.4.2 Efficiency

We compare oTIT with Link-LDA and FLDA in terms of time (500 iterations) and memory efficiency on Sina Weibo dataset from stream $s = 1$ to $s = 6$. Since FLDA-Stream runs independently in each stream, it has nearly the same time and memory cost with oTIT. However, oTIT significantly outperforms FLDA-Stream in terms of precision. Thus, for clarity, we omit the result of FLDA-Stream as well as New-follower. Figure 7 shows the results of efficiency comparison. Due to the step of grid search for λ, both TIT and oTIT cost more time than the others in the first stream. However, in the new coming streams, both time and memory costs of oTIT stay approximately a constant. This is because the costs of oTIT only depend on the size of new streams, while the costs of other methods are accumulative
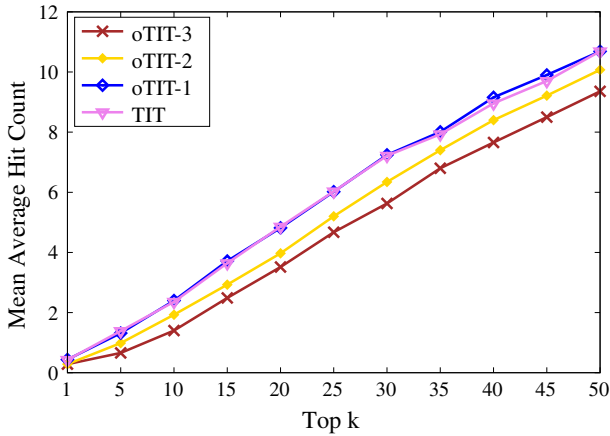
Fig. 6 Mean average hit count for TIT, oTIT-1, oTIT-2 and oTIT-3 across all topics over different settings of k

Fig. 7 Comparisons of different approaches in terms of time (500 iterations) and memory efficiency. **a** Time cost (500 iterations) and **b** memory cost

since they need to scan the whole data repeatedly. It demonstrates that oTIT is more efficient for social influence analysis from large dataset.

4.4.3 Parameter sensitivity

We have determined the optimal value of λ for oTIT through minimizing held-out perplexity on a validation set. However, in the procedure of precision evaluation, we surprisingly find an interesting phenomenon that oTIT can achieve best performance with different settings of λ on different topics. It means the decay rate of influence varies across different topics. Figure 8 shows the results of oTIT over different settings of λ on the topic infant&mom and sport in terms of $HC@50$. We vary λ from 1 to 50 with a step of 1, and we only show the results with respect to $\lambda \in [1, 20]$, due to the stable precision achieved of oTIT when $\lambda > 20$ in each case. Here, smaller values of λ result in faster decay of influence. Figure 8a indicates that oTIT is optimal when $\lambda \geq 15$ on topic sport, which means some users attractive recently are those popular past. On the contrary in Fig. 8b, oTIT is optimal when $\lambda \leq 5$ on topic Infant&Mom, which means some users who was popular past are no longer attractive now. For further understanding this phenomenon, we show the optimal values of λ for oTIT in terms of $HC@50$ on 8 topics in Table 6. We can observe that influence presents different decay patterns on different topics or different kinds of topics. The reasons for this phenomenon may be that, for some topics like food and horoscopes, the influencers are usually not specific individuals but some popular accounts managed by a team. Intuitively, this kind of accounts are prone to lose their attractiveness when other similar accounts become popular. While, for some topics like sport and music, the influencers are usually those famous celebrities whose influence generally can remain for a long time. Further more, the influence of celebrities also has different variation patterns on different topics. For some popular topics like movie and music, users' influence changes relatively fast since the hotspot on these topics keeps shifting. Instead, for some professional topics like finance and education, users' influence tends to remain relatively stable. This valuable discovery reveals a promising feature of social influence that it has different variation patterns according to different topics, which gives us a better understanding of influence and its dynamic nature.

5 Related work

Much effort has been made for social influence analysis and a lot of work has been done on different forms of influence. The correlation between social influence and content similarity was studied in Crandall et al. (2008). Dietz et al. (2007) devised a probabilistic topic model to measure the influence of links between papers. Mehmood et al. (2013) introduced CSI, a information propagation model based on the Independent Cascade model (Kempe et al. 2003), for analyzing the social influence at the granularity of community. Indirected influence is also studied in the work (Liu et al. 2010), where a more than two steps reposting is regarded as the evidence for the exiting of indirected influence. Zhang et al. (2013) studied the locality of social influence from a user's ego network and they found that users' behaviors are mainly influenced by their close friends. Lin et al. (2013) analyzed the external influence of events. They found that the events in social network may be caused by the social influence or by the external influence of the events. Foulds and Smyth (2013) studied the influence of scientific articles. They introduced topical influence that measures to what extent an article tends to spread its topics to the articles that cite it. Romero et al. (2011) introduced a novel influence
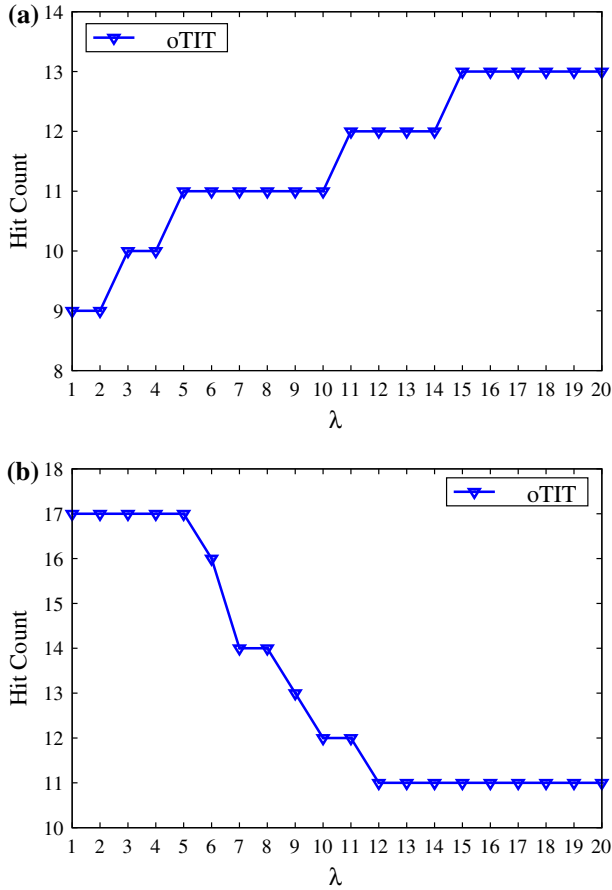
Fig. 8 oTIT over different settings of λ on the topic infant&mom and sport in terms of $HC@50$. **a** Sport category ($HC@50$) and **b** infant&mom category ($HC@50$)

Table 6 Optimal values of λ for oTIT on 8 different topics in terms of $P@(k = 50)$

Topic	Automobile	Horoscopes	Make up	Food
λ	[1, 4]	[1, 4]	[1, 5]	[1, 4]
Topic	Finance	Education	Movie	Music
λ	[11, 20]	[11, 20]	[7, 10]	9

measure that takes into account the passivity of the audience in the social network. They developed a HITS-like iterative algorithm to compute influence.

One of the most significant problems with regard to social influence is the topic-level influence analysis in social network. Nallapati et al. (2011) proposed to study the topic-level influence of documents, which combines ideas from network flow and topic modeling. Cha et al. (2010) showed the fact that users' influence in Twitter varies across different topics and over time, and top influencers hold significant influence over a variety of topics. Weng et al.

(2010) studied the problem of finding topic-sensitive influential Twitterers in twitter. They first constructed link networks among twitterers based on a topic model. Then a PageRank-based method, called TwitterRank, is used to identify the topic-sensitive influencers. Tang et al. (2009) proposed a Topical Factor Graph model, called Topical Affinity Propagation (TAP), for topic-level influence in large-scale networks. Pal and Counts (2011) proposed a probabilistic clustering method using a set of extracted features to produce a ranked list of top authors for a given topic for identifying topical authorities in microblog environments. Lampos et al. (2014) predicted users' topic-related and topic-unrelated impact in Twitter through a regression model. Multiple carefully selected usage statistics features like the proportion of retweets and proportion of non-duplicate tweets are considered in their model. Embar et al. (2015) also extracted several features of influence and used an aggregating influence scores for measuring users' influence. Bi et al. (2014b) proposed a mixture model, called FLDA, that integrates topic discovery and social influence analysis in the same model. Moreover, FLDA introduced a Bernoulli distribution to model the reasons why a user follows another, content-based or content-independent. He also studied the topic-level influence on content sharing services, i.e., Flickr and 500px (Bi et al. 2014a). Unfortunately, all of these works ignore the dynamics of influence and most of them analyze the influence in a static network. Instead, our oTIT model fully utilize the temporal dynamics of influence and works in an online fashion, which enables oTIT more effective and efficient in the task of identifying the current topic-level influencers in the microblog sphere.

Temporal information was used for finding trendsetters who are early adopters that spread new ideas or trends before they become popular (Saez-Trumper et al. 2012). Zhang et al. (2014) proposed Continuous Temporal Dynamic Behavior (ConTyor) to predict the temporal behavior by considering the social influence and personal preference over continuous time. Temporal model was also studied in the topic model for documents. Dynamic topic model (DTM) (Blei and Lafferty 2006) was presented to capture the evolution of topics based on the Markov assumptions over state transitions in time. Instead, Topics over Time (TOT) (Wang and McCallum 2006) parameterized a continuous distribution associated with each topic to capture how the intensity of topics changes over time. Another important topic model is the relational topic model (RTM) (Chang and Blei 2009), which modeled both words and links in a document network. RTM can be used to summarize a network of documents, predict links between them and predict words within them. Gerrish and Blei (2010) studied the influence of document on topics over time. The documents, whose words have higher expected probability in the next time slice, are considered as the influential ones. Different from these works, we model the temporal information about when users interact with each other for identifying the current topic-level influencers on microblogs.

Another interesting problem about influence is the influence maximization problem (Kempe et al. 2003) which is to find k influencers who can trigger a maximized diffusion of information in social networks. Bakshy et al. (2011) studied the influence maximization problem in Twitter where they assumed every user is an influencer. They found that the most cost-efficient influencers are those ordinary users, whose influence is approximately average. Besides these influencers, word-of-mouth information spreads via many small cascades, mostly triggered by ordinary individuals, is also likely to apply generally. Then the information propagation problem was studied ever since the influence maximization work. He et al. (2015) studied the problem of analyzing text-based cascades in social network for tracking the information diffusion. They developed the HawkesTopic model (HTM) by combining Hawkes process and topic model to address this problem. Wang et al. (2015) studied the problem of influence propagation in dynamic network where they modeled the temporal evolution of influence by using hidden Markov chain, from which user relations

can be accurately predicted. Goyal et al. (2010) designed static and time-dependent models for learning the influence probabilities in information diffusion network, since prior works all assumed that the influence probabilities are known a prior in the input network. They pointed out that the dynamics of influence should be considered in the future work. Guille et al. (2013) presented a survey of representative methods dealing with issues of popular topic detection and information diffusion as well as influential spreaders identification. It is worth noting that although the influence maximization problem is similar to our work of influencers identification in that both of them output the top k influencers, they are different essentially. Firstly, influence maximization is proposed special for viral marketing, while the objectives of influencers identification include not only viral marketing but also influencers search (Weng et al. 2010) and opinions gathering (Pal and Counts 2011), etc. Secondly, researchers usually focus on measuring the influence between users, i.e., the ability of user A influence user B, in the influence maximization problem. By contrast, users' absolute influence instead of relative influence is studied in the influencers identification problem. Thirdly, due to different motivations, different kinds of technical routes are utilized to address these two problems. Generally, information diffusion models are devised to address influence maximization problem, while data mining and machine learning approaches are employed for the influencers identification problem.

6 Conclusion

This paper addressed the problem of analyzing the topic-level temporal influence of users for the finding the current influencers on specific topics in microblog sphere. To achieve this, we first propose the topic-level influence over time (TIT) model, a novel probabilistic generative model jointly over text, links and time. Then, we apply an influence decay function to measure the topic-level temporal influence of each user. After that, we combine TIT and influence decay into a united online model (named oTIT), to track the topic-level influencers in social streams. We compare our approach with Link-LDA, FLDA and FLDA-Stream on a real dataset crawled from Sina Weibo. Our qualitative and quantitative evaluations demonstrate the effectiveness and efficiency of our approach. Moreover, we find that users' influence exhibits different variation patterns on different topics, which provides us a new insight into the dynamics of influence. For future work, we plan to design a scalable method to further speed up its implementation.

7 Appendix

7.1 Gibbs sampling derivation

Here we describe the parameter estimation for TIT based on the collapsed Gibbs Samping. The joint distribution of TIT is shown in Eq. 16:

$$
\begin{aligned}
&p(f, t, w, y, x, z | \rho, \gamma, \varepsilon, \tau, \alpha, \beta) \\
&= p(x, z | \alpha) p(w | z, \beta) p(y | \rho) p(f, t | x, \gamma)^y p(f, t | \varepsilon)^{(1-y)}
\end{aligned}
\tag{16}
$$

Firstly, we give the derivation of each part in the joint distribution in Eqs. 17–21. In these derivations, Dirichlet-Multinomial and Beta-Bernoulli conjugates ensure the tractability of the integrals. Specifically, for the first part:

$$
\begin{aligned}
p(x, z | \alpha) &= \int p(x | \theta) p(z | \theta) p(\theta | \alpha) d\theta \\
&= \int \prod_{u=1}^{U} \prod_{k=1}^{K} \theta_{u,k}^{n_{u(w)}^{(k)}} \prod_{u=1}^{U} \prod_{k=1}^{K} \theta_{u,k}^{n_{u(f)}^{(k)}} \prod_{u=1}^{U} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_{u,k}^{\alpha-1} d\theta \\
&= \int \prod_{u=1}^{U} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_{m,k}^{n_{u(w)}^{(k)}+n_{u(f)}^{(k)}+\alpha-1} d\theta \\
&= \prod_{u=1}^{U} \frac{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})}, \\
\overrightarrow{n}_{u(w)} &= \{n_{u(w)}^{(k)}\}_{k=1}^{K}, \ \overrightarrow{n}_{u(f)} = \{n_{u(f)}^{(k)}\}_{k=1}^{K}
\end{aligned}
\tag{17}
$$

For the second part:

$$
\begin{aligned}
p(w | z, \beta) &= \int P(w | z, \varphi) p(\varphi | \beta) d\varphi \\
&= \int \prod_{k=1}^{K} \prod_{w=1}^{W} \varphi_{k,w}^{n_k^{(w)}} \prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{w=1}^{W} \varphi_{k,w}^{\beta-1} d\varphi \\
&= \int \prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{w=1}^{W} \varphi_{k,w}^{n_k^{(w)}+\beta-1} d\varphi \\
&= \prod_{k=1}^{K} \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})}, \ \overrightarrow{n}_k = \left\{ n_k^{(w)} \right\}_{w=1}^{W}
\end{aligned}
\tag{18}
$$

For the third part:

$$
\begin{aligned}
p(y | \rho) &= \int p(y | \mu) p(\mu | \rho) d\mu \\
&= \int \prod_{u=1}^{U} \prod_{y=0}^{1} \mu_{u,y}^{n_u^{(y)}} \prod_{u=1}^{U} \frac{1}{\Delta(\rho)} \prod_{y=0}^{1} \mu_{u,y}^{\rho-1} d\mu \\
&= \int \prod_{u=1}^{U} \frac{1}{\Delta(\rho)} \prod_{y=0}^{1} \mu_{u,y}^{n_u^{(y)}+\rho-1} d\mu \\
&= \prod_{u=1}^{U} \frac{\Delta(\overrightarrow{n}_u + \overrightarrow{\rho})}{\Delta(\overrightarrow{\rho})}, \ \overrightarrow{n}_u = \left\{ n_u^{(y)} \right\}_{y=0}^{1}
\end{aligned}
\tag{19}
$$

For the fourth part:

$$
\begin{aligned}
p(f, t | x, \gamma)^y &= \int p(f, t | x, \sigma) p(\sigma | \gamma) d\sigma \\
&= \int \prod_{k=1}^{K} \prod_{f=1}^{U} \sigma_{k,(f,*)}^{n_k^{(f,*)}} \prod_{k=1}^{K} \frac{1}{\Delta(\gamma)} \prod_{f=1}^{U} \sigma_{k,(f,*)}^{\gamma-1} d\sigma
\end{aligned}
$$

$$= \int \prod_{k=1}^{K} \frac{1}{\Delta(\gamma)} \prod_{f=1}^{U} \sigma_{k,(f,*)}^{n_k^{(f,*)}+\gamma-1} d\sigma$$

$$= \prod_{k=1}^{K} \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\gamma})}{\Delta(\overrightarrow{\gamma})}, \quad \overrightarrow{n}_k = \left\{ n_k^{(f,*)} \right\}_{f=1}^{U} \tag{20}$$

For the fifth part:

$$p(f,t|\varepsilon)^{(1-y)} = \int p(f,t|\pi)^{(1-y)} p(\pi|\varepsilon) d\pi$$

$$= \int \prod_{f=1}^{U} \pi_{(f,*)}^{n_{(f,*)}} \frac{1}{\Delta(\varepsilon)} \prod_{f=1}^{U} \pi_{(f,*)}^{\varepsilon-1} d\pi$$

$$= \int \frac{1}{\Delta(\varepsilon)} \prod_{f=1}^{U} \pi_{(f,*)}^{n_{(f,*)}+\varepsilon-1} d\pi$$

$$= \frac{\Delta(\overrightarrow{n}_{(f,*)} + \overrightarrow{\varepsilon})}{\Delta(\overrightarrow{\varepsilon})}, \quad \overrightarrow{n}_{(f,*)} = \{n_{(f,*)}\}_{f=1}^{U} \tag{21}$$

We use $\Delta(*)$ to denote the normalization factor in the Dirichlet distribution and the Beta distribution for notational convenience. For example in Eq. 20, $\Delta(\overrightarrow{\gamma}) = \int \prod_{f=1}^{U} \sigma_f^{\gamma-1} d\sigma = \frac{\prod_{f=1}^{U} \Gamma(\gamma_f)}{\Gamma(\sum_{f=1}^{U} \gamma_f)}$, where $\Gamma(*)$ is the Gamma function, and likewise in other equations.

Then, by combining Eqs. 17–21, the joint distribution of TIT can be represented as in Eq. 22:

$$p(f,t,w,y,x,z|\rho,\gamma,\varepsilon,\tau,\alpha,\beta)$$

$$= \prod_{u=1}^{U} \frac{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})} \cdot \frac{\Delta(\overrightarrow{n}_u + \overrightarrow{\rho})}{\Delta(\overrightarrow{\rho})}$$

$$\cdot \prod_{k=1}^{K} \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\gamma})}{\Delta(\overrightarrow{\gamma})} \cdot \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})} \cdot \frac{\Delta(\overrightarrow{n}_{(f,*)} + \overrightarrow{\varepsilon})}{\Delta(\overrightarrow{\varepsilon})} \tag{22}$$

After that, we derive the full conditional distributions for z_j with index $j = (u, m)$ in Eq. 23 and x_i, y_i with index $i = (u, l)$ in Eqs. 24 and 25 from the joint distribution. In the following derivations, we use the fact that $\Gamma(x+1) = x \times \Gamma(x)$.

$$p(z_j|z_{\neg j}, x, w, \alpha, \beta) = \frac{p(z, w|\alpha, \beta)}{p(z_{\neg j}, w|\alpha, \beta)}$$

$$\propto \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\beta})}{\Delta(\overrightarrow{n}_{k,\neg j} + \overrightarrow{\beta})} \cdot \frac{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}{\Delta(\overrightarrow{n}_{u(w),\neg j} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}$$

$$= \frac{\Gamma(n_k^{(w)} + \beta) \Gamma\left(\sum_{w=1}^{W} n_{k,\neg j}^{(w)} + W\beta\right)}{\Gamma\left(\sum_{w=1}^{W} n_k^{(w)} + W\beta\right) \Gamma(n_{k,\neg j}^{(w)} + \beta)}$$

$$\cdot \frac{\Gamma(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha)\Gamma\left(\sum_{k=1}^{K}(n_{u(w),\neg j}^{(k)} + n_{u(f)}^{(k)} + \alpha)\right)}{\Gamma(\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha))\Gamma(n_{u(w),\neg j}^{(k)} + n_{u(f)}^{(k)} + \alpha)}$$

$$= \frac{n_{k,\neg j}^{(w)} + \beta}{\sum_{w=1}^{W} n_{k,\neg j}^{(w)} + W\beta} \cdot \frac{n_{u(w),\neg j}^{(k)} + n_{u(f)}^{(k)} + \alpha}{\sum_{k=1}^{K}(n_{u(w),\neg j}^{(k)} + n_{u(f)}^{(k)} + \alpha)}$$

$$\propto \frac{n_{k,\neg j}^{(w)} + \beta}{\sum_{w=1}^{W} n_{k,\neg j}^{(w)} + W\beta}\left(n_{u(w),\neg j}^{(k)} + n_{u(f)}^{(k)} + \alpha\right) \tag{23}$$

$$p(x_i, y_i = 1 | \mathrm{f}, t, x_{\neg i}, y_{\neg i}, z, \alpha, \gamma, \rho) = \frac{p(x, y, f, t, z | \alpha, \gamma, \rho)}{p(x_{\neg i}, y_{\neg i}, f, t, x_{\neg i}, y_{\neg i}, z | \alpha, \gamma, \rho)}$$

$$\propto \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\gamma})}{\Delta(\overrightarrow{n}_{k,\neg i} + \overrightarrow{\gamma})} \cdot \frac{\Delta(n_u^{(y=1)} + \rho_1)}{\Delta(\overrightarrow{n}_{u,\neg i} + \overrightarrow{\rho})} \cdot \frac{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}{\Delta(\overrightarrow{n}_{u(w),\neg i} + \overrightarrow{n}_{u(f),\neg i} + \overrightarrow{\alpha})}$$

$$= \frac{\Gamma\left(\sum_{t=1}^{T} n_k^{(f,t)} + \gamma\right)\Gamma\left(\sum_{f=1}^{U}\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + U\gamma\right)}{\Gamma\left(\sum_{f=1}^{U}\sum_{t=1}^{T} n_k^{(f,t)} + U\gamma\right)\Gamma(\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + \gamma)} \cdot \frac{\Gamma(n_u^{(y=1)} + \rho_1)\Gamma(n_{u,\neg i}^{(y)} + \rho_0 + \rho_1)}{\Gamma(n_u^{(y)} + \rho_0 + \rho_1)\Gamma(n_{u,\neg i}^{(y=1)} + \rho_1)}$$

$$\cdot \frac{\Gamma(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha)\Gamma\left(\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)\right)}{\Gamma\left(\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha)\right)\Gamma(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)}$$

$$= \frac{\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + \gamma}{\sum_{f=1}^{U}\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + U\gamma} \cdot \frac{n_{u,\neg i}^{(y=1)} + \rho_1}{n_{u,\neg i}^{(y)} + \rho_0 + \rho_1} \cdot \frac{n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)}$$

$$\propto \frac{\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + \gamma}{\sum_{f=1}^{U}\sum_{t=1}^{T} n_{k,\neg i}^{(f,t)} + U\gamma}(n_{u,\neg i}^{(y=1)} + \rho_1)(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha) \tag{24}$$

$$p(x_i, y_i = 0 | f, t, x_{\neg i}, y_{\neg i}, z, \alpha, \varepsilon, \rho) = \frac{p(f, t, x, y, z | \alpha, \varepsilon, \rho)}{p(f, t, x_{\neg i}, y_{\neg i}, z | \alpha, \varepsilon, \rho)}$$

$$\propto \frac{\Delta(\overrightarrow{n}_{(f,*)} + \overrightarrow{\varepsilon})}{\Delta(\overrightarrow{n}_{(f,*),\neg i} + \overrightarrow{\varepsilon})} \cdot \frac{\Delta(n_u^{(y=0)} + \rho_0)}{\Delta(\overrightarrow{n}_{u,\neg i} + \overrightarrow{\rho})} \cdot \frac{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f)} + \overrightarrow{\alpha})}{\Delta(\overrightarrow{n}_{u(w)} + \overrightarrow{n}_{u(f),\neg i} + \overrightarrow{\alpha})}$$

$$= \frac{\Gamma\left(\sum_{t=1}^{T} n_{(f,t)} + \varepsilon\right)\Gamma\left(\sum_{f=1}^{U}\sum_{t=1}^{T} n_{(f,t),\neg i} + U\varepsilon\right)}{\Gamma\left(\sum_{f=1}^{U}\sum_{t=1}^{T} n_{(f,t)} + U\varepsilon\right)\Gamma\left(\sum_{t=1}^{T} n_{(f,t),\neg i} + \varepsilon\right)} \cdot \frac{\Gamma(n_u^{(y=0)} + \rho_0)\Gamma(n_{u,\neg i}^{(y)} + \rho_0 + \rho_1)}{\Gamma(n_u^{(y)} + \rho_0 + \rho_1)\Gamma(n_{u,\neg i}^{(y=0)} + \rho_0)}$$

$$\cdot \frac{\Gamma(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha)\Gamma\left(\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)\right)}{\Gamma\left(\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f)}^{(k)} + \alpha)\right)\Gamma(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)}$$

$$= \frac{\sum_{t=1}^{T} n_{(f,t),\neg i} + \varepsilon}{\sum_{f=1}^{U}\sum_{t=1}^{T} n_{(f,t),\neg i} + U\varepsilon} \cdot \frac{n_{u,\neg i}^{(y=0)} + \rho_0}{n_{u,\neg i}^{(y)} + \rho_0 + \rho_1} \cdot \frac{n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K}(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha)}$$

$$\propto \frac{\sum_{t=1}^{T} n_{(f,t),\neg i} + \varepsilon}{\sum_{f=1}^{U}\sum_{t=1}^{T} n_{(f,t),\neg i} + U\varepsilon}(n_{u,\neg i}^{(y=0)} + \rho_0)(n_{u(w)}^{(k)} + n_{u(f),\neg i}^{(k)} + \alpha) \tag{25}$$

Finally, the latent parameters can be estimated by the assigned samples, which has been given in Eqs. 4–8.

7.2 Key topics

Among 36 categories in ground truth, 29 of them are well clustered in the experiment. Here, we show the manually labeled topics with top keywords achieved by oTIT in Table 7.

Table 7 29 Topics achieved by oTIT

Topic	Keywords
Infant&mom	Infant, mom, dad, baby, parent, child, love, pregnancy
Sport	Game, game fans, football, score, player, team, sport, basketball
Anime	Anime, post production, picture, comics, cartoon, production, update
Medical treatment	Doctor, hospital, cure, patient, surgery, health, recover, medical treatment
Finance	Investment, stock market, economics, market, fund, capital, Shanghai composite index
Pet	Cat, dog, owner, animal, pet, adoption, stray, home, cute
Music	Music, band, music festival, song, rock, show, album, guitar
Cate	Cate, restaurant, flavor, shop, foodie, delicacy, hotpot, cuisine , seafood, liquor
Automobile	Automobile, vehicle, new, Audi, Benz, engine, Volkswagen, drive, power
Movie	Movie, story, online, volume, release, documentary, starring, douban, America
TV plays	Actor, TV plays, director, satellite television, film and television, shoot, scriptwriter, starring
Law	Lawyer, law, court, news, government, investigate, journalist, report, case, officer
Fitness	Action, thin, bodybuilding, leg, exercise, train, lose weight, muscle, yoga
Education	Teacher, student, examination, English, study, university, school, education, curriculum
Photography	Photography, photographer, shoot, photo, work, camera, camera lens, picture
Regimen	Health, food, body, vitamin, nutrition, function, regimen, body, diet
Read	Book, read, story, write, publish, fiction, writer, literature, reader, new book
Game	Game, play, game player, live show, hero, alliances, the Three Kingdoms
Horoscopes	Horoscopes, Virgo, Leo, Scorpio, Libra, Sagittarius, Scorpio, Aquarius, Cancer, Taurus, Pisces
Science	Research, human, science, nation, scientist, earth, event, protect

Table 7 continued	Topic	Keywords
	Tour	Tour, travel, Tibet, scenic spot, beautiful, Lijiang, scenery, Yunnan, park, tourist
	Make up	Skin, mask, effect, product, moisturize, essence, skin whitening, skin-revitalizing, Korea,
	Fashion	Fashion, style, brand, model, stylist, magazine, design, costume, hairstyle, fashionable dress
	Internet	Company, enterprise, Internet, start up business, marketing, industry, brand, advertizement
	Music	Music, band, music festival, song, rock, show, album, guitar
	Emotion	Love, time, matter, sometimes, forever, feeling, emotion, understand
	Collection	Tea, culture, tradition, museum, calligraphy, history, character, collection, ancient
	Art	Art, artist, create, work, exhibition, gallery, painter, France, culture
	Real estate	City, government, million, real estate, house price, bank, billion, money, income

References

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Eighth IEEE international conference on data mining, 2008. ICDM'08* (pp. 3–12). IEEE.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65–74). ACM.

Bi, B., Kao, C., Wan, C., & Cho, J. (2014a). Who are experts specializing in landscape photography? In *20th ACM SIGKDD conference on knowledge discovery and data mining*.

Bi, B., Tian, Y., Sismanis, Y., Balmin, A., Cho, J. (2014b). Scalable topic-specific influence analysis on microblogs. In *Proceedings of the 7th ACM international conference on Web search and data mining* (pp. 513–522). ACM.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, *10*(10–17), 30.

Chang, J., & Blei, D. M. (2009). Relational topic models for document networks. *AIStats*, *9*, 81–88.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 160–168). ACM.

Diao, Q., Jiang, J., Zhu, F., & Lim, E. P. (2012). Finding bursty topics from microblogs. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers* (Vol. 1, pp. 536–544). Association for Computational Linguistics.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning* (pp. 233–240). ACM.

Embar, V. R., Bhattacharya, I., Pandit, V., & Vaculin, R. (2015). Online topic-based social influence analysis for the wimbledon championships. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1759–1768). ACM.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5220–5227.

Foulds, J. R., & Smyth, P. (2013). Modeling scientific impact with topical influence regression. In *EMNLP* (pp. 113–123).

Gerrish, S., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. *ICML*, *10*, 375–382.

Goyal, A., Bonchi, F., & Lakshmanan, L. V.(2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 241–250). ACM.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, *42*(2), 17–28.

He, X., Rekatsinas, T., Foulds, J., Getoor, L., & Liu, Y. (2015). Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *Proceedings of the 32nd international conference on machine learning*.

Kempe, D., Kleinberg, J., Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 137–146). ACM.

Lampos, V., Aletras, N., & Preoţiuc-Pietro, D., & Cohn, T. (2014). Predicting and characterising user impact on twitter. In *14th conference of the European chapter of the Association for Computational Linguistics 2014, EACL 2014* (pp. 405–413).

Lin, S., Wang, F., Hu, Q., & Yu, P. S. (2013). Extracting social events for learning better information diffusion models. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 365–373). ACM.

Liu, L., Tang, J., Han, J., Jiang, M., & Yang, S. (2010). Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 199–208). ACM.

Mehmood, Y., Barbieri, N., Bonchi, F., & Ukkonen, A. (2013). CSI: Community-level social influence analysis. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine learning and knowledge discovery in databases* (pp. 48–63). Springer.

Nallapati, R., Mcfarland, D. A., & Manning, C. D. (2011). Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents. In *AISTATS* (pp. 543–551).

Pal, A., & Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 45–54). ACM.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (pp. 18–33). Springer.

Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., & Benevenuto, F. (2012). Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1014–1022). ACM.

Subbian, K., Aggarwal, C. C., & Srivastava, J. (2016). Querying and tracking influencers in social streams. In *Proceedings of the ninth ACM international conference on Web search and data mining* (pp. 493–502). ACM.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 807–816). ACM.

Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L., et al. (2011). Topic level expertise search over heterogeneous networks. *Machine Learning*, *82*(2), 211–237.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433). ACM.

Wang, Z., Wang, S., & Huang, Q. (2015). Latent influence propagation on dynamic networks. In *2015 IEEE China summit and international conference on signal and information processing (ChinaSIP)* (pp. 777–781). IEEE.

Wang, Y., Zhang, Z., Su, S., Chang, C., & Zia, M. A. (2016). Topic-level influencers identification in the microblog sphere. In *ECAI* (pp. 1559–1560).

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261–270). ACM.

Zhang, J., Liu, B., Tang, J., Chen, T., & Li, J. (2013). Social influence locality for modeling retweeting behaviors. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2761–2767). AAAI Press.

Zhang, J., Wang, C., Wang, J., & Yu, J. X. (2014). Inferring continuous dynamic social influence and personal preference for temporal behavior prediction. *Proceedings of the VLDB Endowment*, *8*(3), 269–280.