Inferring the Diffusion and Evolution of Topics in Social Communities

Cindy Xide Lin[1] Qiaozhu Mei[2] Yunliang Jiang[1] Jiawei Han[1] Shanxiang Qi[1]

*[1]Department of Computer Science, University of Illinois at Urbana-Champaign,
201 N. Goodwin Ave., Urbana, IL 61801, USA
[2]School of Information, University of Michigan,
1085 S. University Ave., Ann Arbor, MI 48109, USA
xidelin2@uiuc.edu, qmei@umich.edu, jiang8@uiuc.edu, hanj@cs.uiuc.edu, sqi2@uiuc.edu

ABSTRACT

The prevailing of Web 2.0 techniques has led to the boom of various online communities, where topics are spreading ubiquitously among user-generated documents. Together with this diffusion process is the content evolution of the topics, where novel contents are introduced in by documents which adopt the topic. Unlike an explicit user behavior (*e.g.*, buying a DVD), both the diffusion paths and the evolutionary process of a topic are implicit, making them much more challenging to be discovered.

In this paper, we aim to simultaneously track the evolution of any arbitrary topic and reveal the latent diffusion paths of that topic in a social community. A novel and principled probabilistic model is proposed which casts our task as an joint inference problem, taking into consideration of textual documents, social influences, and topic evolution in a unified way. Specifically, a mixture model is introduced to model the generation of text according to the diffusion and the evolution of the topic, while the whole diffusion process is regularized with user-level social influences through a Gaussian Markov Random Field.

Experiments on both synthetic data and real world data show that the discovery of topic diffusion and evolution benefits from this joint inference; and the probabilistic model we propose performs significantly better than existing methods.

Categories and Subject Descriptors: H.2.8 [**Information Systems Applications**]: Database Applications – *data mining*.

General Terms: Algorithm, Experimentation

Keywords: Information diffusion and evolution, Social networks, Topic modeling, Gaussian markov random field

1. INTRODUCTION

The prevailing of Web 2.0 techniques has led to the boom of various online communities. One of the core problems in analyzing such online communities is concerned with understanding the cascading behaviors and the diffusion of information. Epidemic diseases, adoption of innovation, memes of information, and many types of user actions all spread widely in these communities, following the social network of users. The modeling of information diffusion plays a crucial role in many domains. The contagion of disease forms the foundation of epidemics; the social influence in cascading behaviors has been a basic mechanism of viral marketing; and the diffusion of topics is essential to the understanding of scientific innovation. Recently, a large body of research work has been done in the field of social network analysis, aiming to describe the macro-level dynamics and characteristics of information diffusion [21, 13], reveal key factors that affect the adoption of behaviors [1, 22], and design contagion models that simulate the diffusion process [34, 47].

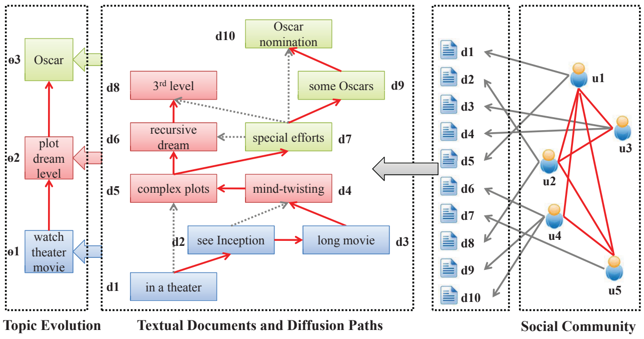


Figure 1: Example of Topic Diffusion and Evolution

Many conclusions in this line of work are motivated and validated in scenarios where the actual contagion/diffusion paths are observed. Such an assumption, which is considered as a common practice in user surveys and controlled user studies, does not apply to large scale online communities however. While the adoptions of behaviors are relatively easy to observe (based on which most macro-level descriptive statistics are computed), the evidence

*This work was supported in part by the U.S. NSF grant IIS-09-05215 and by the Army Research Laboratory accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation stated here on.

The 5th SNA-KDD Workshop '11 August 21, 2011, San Diego, CA, USA.
Copyright 2011 ACM 978-1-4503-0225-8 ...$5.00.

of actual contagion and influence tend to be vague. Who infected whom? Who got the gossip from whom? Who influenced whose research? There are still substantial challenges in this micro-level analysis of information diffusion in large scale social networks. Indeed, users who joined a community or purchased an iPad usually won't explain which particular friends have influenced them; rumor spreaders tend to cover the source of the information; a researcher cites many references in her paper, without labeling the top three that have the most salient influence on her work. The identification of contagion is difficult even if the general social network structure is observed. It is a non-trivial task to detect the actual diffusion paths of user behaviors merely based on the time of adoption and the social network structure, known as the problem of diffusion (or influence) inference [9].

Inference of diffusion becomes even more challenging when the behaviors themselves are subtle. The adoption of explicit behaviors can be easily identified, for instance buying a DVD, joining a community, or using a hashtag in a tweet. Some behaviors are however implicit, such as writing about of a topic, holding an opinion, or having a particular mood. In this paper, we focus on the diffusion of topics in social communities. Inferring topic diffusion has introduced several additional challenges on top of the diffusion inference of explicit behaviors. First, topics are implicit and abstract concepts used in natural language. The adoption of topics cannot be directly identified, instead has to be inferred from the user-generated contents. Second, the meaning of a topic is evolving over time. A smart system should understand that *'MSN search'*, *'Live search'*, and *'Bing'* all refer to the same topic *'the Microsoft search engine'*, with unique aspects at different time; and it should be able to track and adapt to this content change in the the inference of topic diffusion. Third, information transmission is a complex social-psychological behavior [28], so the diffusion process of contents is inevitably influenced by the social relationships of the users. Moreover, the evolution and the diffusion of topics are compound processes: indeed, when a topic spreads from one user to another, new perspectives or new focus is introduced to the topic; and an outbreak of a topic is usually accompanied by a shift of the meaning of the topic. Although there has been a line of work on the diffusion inference of explicit behaviors recently [6, 11, 12, 20, 44, 15, 34, 16, 17, 6, 13, 1, 24, 22], none of this work addresses these challenges, making the existing methods incapable to accurately infer the diffusion paths of topics.

In this paper, we address these challenges by studying the joint inference of topic diffusion and evolution in social communities. Content and linkage in user-generated text information, together with social network structures, are used to facilitate the identification of topic adoption, the tracking of topic evolution, and the estimation of actual diffusion paths of any arbitrary topic. Our intuition is illustrated in Figure 1.

When a topic is introduced into the community by a user, other users read the documents she wrote (*e.g.*, tweets, blogs, scientific papers, *etc*) and adopt the topic by writing about it in their own articles. They may or may not cite the original document, or they may cite it together with other documents. Although topics are spread among documents instead of through social connections, we consider it is much more likely for users to adopt ideas from their social connections (*e.g.*, friends, people they follow, or people they have cited before) than from a stranger. Each document can not only adopt content from documents that influenced it, but also include novel perspectives into the topic, and pass on the *'innovation'* to other documents. The meaning of the topic is thus evolving over time. The goal of the joint inference of topic diffusion and topic evolution is to identify the *'real'* paths through which the topic propagates (red edges among documents in Figure 1), and also identify the time specific versions of the topic.

In this paper, we propose a novel statistical model for topic-based information diffusion and evolution (TIDE). Specifically, a mixture model is introduced to model the generation of text according to the diffusion and the evolution of the topic, while the whole diffusion process is regularized with user-level social influences through a Gaussian Markov Random Field. The discovery of novel aspects and the diffusion paths of the topic can be done by the joint inference of topic diffusion and evolution in TIDE.

Organization. The rest of this paper is organized as follows: Section 2 formally defines the problem of TIDE, as the solution of which a statistical model is proposed in Section 3. We present experiments and results in Section 4, discuss the related work in Section 5, and conclude in Section 6.

2. PROBLEM FORMULATION

In this section, we formally define the task of inferring the diffusion process and tracking the evolution of topics in social communities. We begin with a few key concepts as follows.

Definition 2.1. **Social Network**. A *network* is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of edges among V. Particularly in a *social network*, a vertex corresponds to a user, and an edge $e = (i, j)$ stands for a connection (or a tie) between two users i and j. The strength of the tie (i, j) is defined as a non-negative value $g(i, j)$. An edge can be either *directed* or *undirected*.

Definition 2.2. **Document Collection**. A textual document d_i in a *document collection* $\mathcal{D} = \{d_i\}_{i=1}^{M}$ is defined as a bag of words from a fixed vocabulary $\mathcal{W} = \{w_k\}_{k=1}^{L}$. That is, $d_i = \{c(d_i, w_k)\}_{k=1}^{L}$, where $c(d, w)$ denotes the number of occurrences of word w in d.

Definition 2.3. **Social Community**. A *social community* is defined as the union of a social network $\mathcal{G}$ and a user-generated document collection $\mathcal{D}$, saying $\{\mathcal{G}, \mathcal{D}\}$. Each document $d_i \in \mathcal{D}$ is associated with an *author* a_i in $\mathcal{G}$ and a *time-stamp*s $t_i \in 1..T$.

Definition 2.4. **Topic**. A semantic *topic* θ observed in a particular time period is defined as a multinomial distribution of words $\{p(w|\theta)\}_{w\in\mathcal{W}}$ with the constraint $\sum_{w\in\mathcal{W}} p(w|\theta) = 1$.

Definition 2.5. **Theme**. We define a general and coherent *theme* discussed in a social community as a stream of time-stamped topics $\Theta = \{\theta_t\}_{t=0}^{T}$. We call θ_0 the *primitive topic*, which represents the original content of the theme prior to the discussions of the social community. $\theta_{t>0}$ are time variant versions of θ_0, which are gradually developed in the discussions of the social community. That is, θ_t is the snapshot of Θ at time t, which represents the novel aspect of the theme appearing at time t. Altogether Θ represents the origin and evolution of the contents of the theme over time.

While the text content of individual document can be explicitly observed, the general semantics of the time-variant topics and the adoption of the topic(s) in a document is implicit. What else remains implicit is the source adopted in a document. There could naturally be multiple sources: a document can be influenced by a few other documents, thus inherit the topic from those documents. The influence of some sources can be more salient than the influence of others. A document could also introduce original perspectives of the topic without being influenced by any existing document. The existence and strengths of the influence among documents assemble the actual diffusion process of the topic, which is formally defined as a *diffusion graph*.

Definition 2.6. **Diffusion Graph**. Given a theme Θ, we define a *diffusion flow* from one document d_j to another d_i ($t_j < t_i$) as the likelihood that d_i adopted the topic of Θ due to the influence of d_j. The strength of such a diffusion flow is denoted as a positive value $\pi_{i,j}$. Note that d_i can also introduces its novel perspective to Θ. In this case, we assume there is a diffusion flow into d_i from the time-stamped topic θ_{t_i}, with a strength $\pi_{i,\theta}$. Therefore, we define *diffusion vector* π_i as a vector of the strength of all the diffusion flows into d_i, *i.e.*, $\pi(i) = \{\pi_{i,j}\}_{d_j \in \mathcal{D}} \cup \{\pi_{i,\theta}\}$, with the constraint $\sum_{d_j \in \mathcal{D}} \pi_{i,j} + \pi_{i,\theta} = 1$. The union of diffusion flows into all documents in $\mathcal{D}$ assembles the *diffusion graph*, *i.e.*, $\Pi = \{\pi(i)\}_{d_i \in \mathcal{D}}$. Clearly, the graph Π is both weighted and directed.

Although the actual diffusion graph is unobserved, there are proxy networks that convey weaker signals in social communities. In many scenarios, a reference network (denoted as $\mathcal{R}$) of the documents can be observed, for example the citation network of scientific publications, the hyperlink network of blog articles, or the tweet network posted by follower-followees. Intuitively, the diffusion network should correlate well with such a reference network because a document is very likely to be influenced by documents it cites. However, the actual diffusion network could still be substantially different from $\mathcal{R}$, because many real influential references are covered up, and many explicitly cited ones do not represent the true influence. Another signal is the social network structure. An author is likely to follow the work of his social connections, such likely to adopt topics and ideas from the documents they generate [30, 40]. We call the set of documents pointing to d_i in the reference network as d_i's *reference set*, denoted as $r(i) \subset \mathcal{D}$. When no signal of citation or social communication is available, r_i can be simply instantiated as all documents with a time stampe prior to t_i. When such a reference network is available, we assume $\pi_{i,j} = 0$ if $j \notin r(i)$. Clearly, we also have $\pi_{i,i} = 0$.

Based on the definitions of concepts above, we can formalize the two major tasks of tracking **the diffusion and evolution of topics in social communities**. Given the input of a social community $\mathcal{G}$, a user-generated document collection $\mathcal{D}$, and the primitive topic θ_0 defining a theme, we aim to:

Task 1: Infer the Diffusion Graph. In this task, the goal is to discover the latent diffusion flow graph documents (and topics) (*i.e.* Π). The result of this task can be used to answer (i) *the source(s) of topic in a document*: to what extent the document is influenced by other documents, and (ii) *the degree of originality in a document*: how much novel perspetives the document introduces to the topic.

Task 2: Track Topic Evolution. In this task, the goal is to infer the time-variant versions of topics (*i.e.*, $\{\theta_t\}_{t=1}^{T}$) of a theme. By inferring Θ given θ_0, we expect to keep track of the new developments of the theme, understand its evolution over time, and better understand how it influences documents, *etc.*

The two tasks are challenging in many ways. First, although recently there are extensive studies on inferring the social influence on explicit behaviors at the user level [6, 42, 12, 11, 45], there is limited progress in the analysis of the influence on the adoption of latent topics at document level [20, 44]. Even though topic diffusion occurs at document level, the influence along the social network structure is playing an non-negligible role. There is however little existing wisdom on how to bridge document networks with social networks. The implicity nature of topics have made the problem even harder. Second, the inference of topic diffusion cannot be done independently to the tracking of topic evolution. Along with the diffusion process of the topic, new contents are introduced into the topic, making the semantics of the topic evolving over time. Without understanding the shift of topic contents, it is impossible to accurately detect the adoption of the topic in documents especially after a substantially long time. Moreover, since usually there are limited labeled examples, the solution model should be unsupervised. All of these challenges require us to propose a unified model that takes social connections, textual contents, influence among documents, and temporal information into consideration. In the following section, we propose such an integrative model and present the joint inference of topic diffusion and evolution.

3. PROPOSED MODELS

In this section, we propose a novel and integrative probabilistic model of Text-based Information Diffusion and Evolution (TIDE) in social communities. Based on TIDE, we present the joint infernce of the diffusion graph and the evolution of arbitrary topics.

3.1 Intuitions and the General Model

The general model of TIDE is designed based on a few key observations in social communities.

Observation 1. *Diffusion and Contents.* When there is a significant diffusive flow between two documents, or there is a significant influence on one document on the other, the content of these two documents tend to be highly related. On the other hand, if two documents talk about different subjects, there is unlikely salient influence or significant diffusion flow between them even if one cites the other [36]. *W.l.o.g.*, we can assume that the content of a document depends on the documents which have influenced it.

Observation 2. *Diffusion and Social Connections.* Information transmission is a complex social-psychological behavior [28], *e.g.*, there exist persistency interests of users [32]. The diffusion process among documents is likely to be regularized by social connections of their authors. Indeed, an author is more likely to follow the work of her friends and thus adopt topics and ideas from a friend instead of from a random author. The diffusion flows among documents are thus dependent to the social network of authors.

Observation 3. *Diffusion and Evolution.* As the diffusion proceeds, both the semantics of the topic and the regularization effect of the social network of users evolve over time. If an aspect in a document never appear in any of its potential references (either papers it cites or all existing papers exposed to its author), it is likely to be original ideas introduced by the document, which contributes to the evolution of the general theme. Meanwhile, the strength of influence through old social connections would decay after a reasonably long time.

Given a collection of authored and time-stamped documents $\mathcal{D}$, a social community $\mathcal{G}$ of users who published these documents, and a primitive topic θ_0 representing the original semantics of a theme, we aim at inferring the latent stream of topics Θ and the diffusion graph Π. Based on our observations above, the task of TIDE is then cast as the joint inference of the posterior of Θ and Π:

Formally, our object becomes to infer:

$$P(\Pi, \Theta | \mathcal{G}, \mathcal{D}, \theta_0) \propto P(\Theta | \Pi, \mathcal{D}, \theta_0) \cdot P(\Pi | \mathcal{G}) \qquad (1)$$

Based on our observations, here we assume that the generation of the diffusion graph (only) depends on the social network structure, while the evolution of topics depends on the documents, the diffusion process, and of course the original version of the topic. We denote the first component of Equation 1 as the *topic model* and the second as the *diffusion model*. Please note that although TIDE can be easily extended to model the mixture of multiple topics (similar to LDA [3]), we only present the primitive case to model only one given topic. Our focus is to model the diffusion and evolution of any given topic instead of the discovery of multiple topics. We leave the modeling of multiple topics in our future work.

In the topic model, a *mixture model* is designed to extract the topic snapshots (time-variant versions) of the theme (Section 3.2). In the diffusion model, we introduce a *Gaussian markov random field* based on *graph projection* to model the dependency of diffusion flows on social connections (Section 3.3). Finally, the inference of the combined model is discussed in Section 3.4.

3.2 The Topic Model

It is difficult to directly compute the posterior of topics Θ. We make the following transformation such that

$$P(\Theta|\Pi, \mathcal{D}, \theta_0) \propto P(\mathcal{D}|\Theta, \Pi, \theta_0) \cdot P(\Theta|\theta_0), \tag{2}$$

where the introduction of new aspects to the topic (*i.e.*, the time-variant topic snapshots) does not depend on the diffusion flows.

We consider a typical generative process of $\mathcal{D}$: each document d_i is generated from a mixture model. When writing each word in d_i, one first chooses a component model from the mixture with a certain probability; once the component model θ is selected, a word is sampled according to the word distribution of θ.

We first introduce a background component model θ_B estimated from the entire collection that explains the generation of common English words in the document d_i. The rest component models are designed based on the diffusion flows. Specifically, we introduce a component model for each document d_j that could have potentially influenced d_i. There is a non-trivial diffusion flow from d_j to d_i, and d_i could inherit the topic of d_j according to the strength of this diffusion. These component models can be estimated simply using a maximum likelihood estimator on the corresponding d_j. Finally, we introduce a component model to explain the novel aspects introduced by the document d_i, *i.e.*, the aspect that is not influenced by any existing document. We assume that this aspect is generated directly from the latent topic at the time that d_i is written (θ_{t_i}). In other words, the original content is diffused from the topic directly to the document instead of from other documents. We assume that the probability of choosing each component is proportional to the strength of the diffusion vector, *i.e.*, $\pi(i)$.

Formally, the probability of generating a word w in d_i is:

$$\begin{aligned} & p(w|d_i) \\ & = (1 - \lambda_B)(\sum_{j \in r(i)} \pi_{i,j} p(w|\theta_{d_j}) + \pi_{i,\theta} p(w|\theta_{t_i})) + \lambda_B p(w|\theta_B) \end{aligned}$$

where λ_B is a predefined parameter that fixes the sampling probability of the background model. Note that for documents $d_j \notin r(i)$, we have $\pi_{i,j} = 0$. The likelihood of the collection $\mathcal{D}$ is given as:

$$P(\mathcal{D}|\Pi, \Theta, \theta_0) = \prod_{d_i \in \mathcal{D}} \prod_{w \in \mathcal{W}} p(w|d_i)^{c(w,d_i)}$$

We then consider the generation of the time-variant versions of the topic, Θ. In TIDE, the primitive topic θ_0 is realized as a conjugate Dirichlet prior of the time-variant topic model θ_t: $Dir(\{1 + \mu_E p(w|\theta_0)\}_{w \in \mathcal{W}})$. By doing so, we regularize these time-variant topic snapshots so that they can reflect the novel aspects of the theme, but do not shift away from it. μ_E indicates how much we rely on the prior. Formally,

$$P(\Theta|\Pi) = \prod_{t \in 1..T} p(\theta_t|\theta_0) = \prod_{t \in 1..T} \prod_{w \in \mathcal{W}} p(w|\theta_t)^{\mu_E p(w|\theta_0)}$$

3.3 The Diffusion Model

Comparing to the modeling of topic evolution, the modeling of diffusion graph ($P(\Pi|\mathcal{G})$) is less straightforward. Intuitively, the diffusion graph Π should be regularized by the social network $\mathcal{G}$, as social influence plays an important role in topic diffusion. However, Π is a network of *documents* while $\mathcal{G}$ is a network of users. This makes it hard to model the regulation effect of $\mathcal{G}$ on Π. We need a bridge between the two heterogenous networks, for which we introduce the operation of *graph projection*.

Definition 3.1. **Graph Projection**. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two graphs, a projection $f : \mathcal{G}_1|\mathcal{G}_2 \to \mathcal{G}'_1$ is called a *graph projection* if:

1. $\mathcal{V}(\mathcal{G}'_1) = \mathcal{V}(\mathcal{G}_2)$.
2. $\forall v \in \mathcal{V}(\mathcal{G}'_1), \exists u \in \mathcal{V}(\mathcal{G}_1)\ s.t.\ v \in f(u)$.
3. $\forall e = (u, v) \in \mathcal{E}(\mathcal{G}'_1), \forall u' \in f(u)\ and\ v' \in f(v), e' = (u', v') \in \mathcal{E}(\mathcal{G}'_1)$.

Through graph projection, two networks are endowed with the same vertex set, so that the comparison of them becomes more succinct and natural. Note that there are two asymmetric projection directions: 1) projecting $\mathcal{G}$ into a document network and using it as *a priori* of Π, or 2) projecting Π into a social network and consider the generation of such a social network based on $\mathcal{G}$. Since the document collection $\mathcal{D}$ is commonly much larger than the set of user $\mathcal{V}(\mathcal{G})$, projecting the document network into a social network is at inevitable risk of losing information. Although this doesn't rule out the second direction of graph projection, in this work we consider the first direction: the projection of $\mathcal{G}$ into a document network.

Let's denote Π' as the document network projected from $\mathcal{G}$, s.t.

$$P(\Pi|\mathcal{G}) = P(\Pi|\Pi') = P(\{\pi(i)\}_{d_i \in \mathcal{D}}|\Pi').$$

The remaining issue is how to fold the $\mathcal{G}$ into Π' and how to model the generation of Π based on Π'. Note that like Π, we can also denote $\Pi' = \pi'(i)_{d_i \in \mathcal{D}}$. We start with the generative model $P(\Pi|\Pi')$.

Gaussian Graphicl Models (GGM) [46] are classical models used to explain the generation of networks, which could be an ideal solution of our problem. In a typical GGM, each nodes in the graph is modeled as a random variable, for example a vector of k features. In our scenario, such a vector can be implemented as the diffusion vector $\pi(i)$. The joint distribution of all these variables (in our case, $P(\pi(i))$) is assumed to be a multivariate Gaussian. Each edge in Π' stands for the conditional dependency between two Gaussian variables, thus the graph structure Π' corresponds to the inverse covariance matrix.

However, the computational complexity of such a graphical model usually scales cubically with the number of variables, and therefore becomes intolerant even for a moderate size of dataset. To make our model practical, we introduce an independency assumption: the diffusion vector of one document is independent to the others. By doing so, we can simplify the generative model of Π as

$$P(\Pi|\Pi') = \prod_{d_i \in \mathcal{D}} P(\pi(i)|\pi'(i)) \tag{3}$$

Here $\pi'(i) = \{\pi'_{i,j}\}_{j \in r(i)} \cup \{\pi'_{i,\theta}\}$ is a conjugate prior vector, indicating the expected value of $\pi(i)$. Since Π' is projected from $\mathcal{G}$, $\pi'_{i,j}$ represents the social influence between a_j (the author of d_j) to a_i, which decays over time. By doing this, the document-level influence is regulated by the social tie at the user level.

Formally, we define $\pi'_{i,j} = \frac{1}{Z(\pi'(i))} g(a_i, a_j) \cdot e^{-\frac{t_i - t_j}{\alpha}}$ by consolidating an exponential time model with $\mathcal{G}$ [1]. Intuitively, doc-

[1] Other decay functions are also applicable [7].

uments with higher authority is likely to introduce more original content. We thus define $\pi'_{i,\theta} = \frac{1}{Z(\pi'(i))} Aut(a_i)$, where $Aut(a_i)$ is an estimation of the authority of d_i. $Z(\pi'(i))$ is a normalization factor such that $\sum_{d_j \in \mathcal{D}} \pi'_{i,j} + \pi'_{i,\theta} = 1$.

Given the design of $\pi' i)$, the computation of $P(\pi(i)|\pi'(i))$ is still non-trivial because of the dependency between the dimensions of $\pi(i)$. We introduce a *Gaussian Markov Random Field* [33] to model the conditional probability $P(\pi(i)|\pi'(i))$ for each d_i.

Definition 3.2. **Gaussian Markov Random Field (GMRF)** . A random vector $\S = (x_1, x_2, \cdots, x_n)^T$ is called a GMRF *w.r.t.*the graph $\mathcal{G} = (\mathcal{V} = \{1, 2, \cdots, n\}, \mathcal{E})$ with the mean μ and the precision matrix $\mathcal{Q}_\S$, iff the density of $\S$ has the form

$$P(\S) = (2\pi)^{-n/2} |\mathcal{Q}_\S|^{1/2} e^{-\frac{1}{2}(\S-\mu)^T \mathcal{Q}_\S (\S-\mu)}$$

and $\mathcal{Q}_\S(i,j) \neq 0 \Leftrightarrow (i,j) \in \mathcal{E}$ for all $i \neq j$.

In our case, the random vector is the diffusion vector $\pi(i)$, with the mean as the prior vector $\pi'(i)$. The precision matrix $\mathcal{Q}_{\pi(i)}$ corresponds to the similarities between the dimensions of $\pi(i)$ (documents and topic snapshots), which can be realized as the content similarities of corresponding θ_{d_j}'s and θ_t's. Computationally, $P(\pi(i)|\pi'(i))$ is defined as:

$$\begin{aligned} &P(\pi(i)|\pi'(i)) \\ &\propto e^{-\frac{1}{2} \sum\limits_{i',j' \in \{r(i)\} \cup \{\theta\}} (\pi_{i,i'} - \mu_{i,i'}) \mathcal{Q}_{\pi(i)}(i',j')(\pi_{i,j'} - \mu_{i,j'})} \end{aligned}$$

3.4 Parameter Estimation

Given our model defined above, we can fit the model to the data and estimate the parameters using a Maximum A Posterior estimator [38]. Expectation Maximization (EM) algorithm [29] is applied, which iteratively computes a local maximum of the posterior. Computationally, the log likelihood we want to maximize is:

$$\begin{aligned} &E_{\Lambda^{(n-1)}}\{\log p(C|\Lambda)p(\Lambda)\} \propto \qquad (4) \\ &\sum_{d_i, w, d_j \in r(i)} c(d_i, w)(1 - z^{(n)}_{d_i,w}(\theta_B)) z^{(n)}_{d_i,w}(\theta_{d_j}) \log((1-\lambda_B)\pi_{i,j} p(w|\theta_{d_j})) \\ &+ \sum_{d_i,w} c(d_i, w)(1 - z^{(n)}_{d_i,w}(\theta_B)) z^{(n)}_{d_i,w}(\theta_{t_i}) \log((1-\lambda_B)\pi_{i,E} p(w|\theta_{t_i})) \\ &+ \sum_{d_i,w} c(d_i, w) z^{(n)}_{d_i,w}(\theta_B) \log(\lambda_B p(w|\theta_B)) + \mu_E \sum_{\theta_t, w} p(w|\theta_0) \log p(w|\theta_t) \\ &- \frac{\mu_G}{2} \sum_{d_i} \sum_{i',j' \in \mathcal{N}(i)} (\pi_{i,i'} - \mu_{i,i'}) \mathcal{Q}_{pi(i)}(i',j')(\pi_{i,j'} - \mu_{i,j'}) \end{aligned}$$

Here μ_G is a weight combining two components, and we use terms $z_{d_i,w}(\cdot)$ instead of $p(z_{d_i,w} = \cdot)$ for better equation display.

In the E-Step, we compute the expectation of the hidden variables:

$$z^{(n)}_{d_i,w}(\theta_{d_j}) = \frac{\pi^{(n-1)}_{i,j} p(w|\theta_{d_j})}{\sum\limits_{j' \in r(i)} \pi^{(n-1)}_{i,j'} p(w|\theta_{d_{j'}}) + \pi^{(n-1)}_{i,\theta} p(w|\theta_{t_i})}$$

$$z^{(n)}_{d_i,w}(\theta_{t_i}) = \frac{\pi^{(n-1)}_{i,\theta} p^{(n-1)}(w|\theta_{t_i})}{\sum\limits_{j' \in r(i)} \pi^{(n-1)}_{i,j'} p(w|\theta_{d_{j'}}) + \pi^{(n-1)}_{i,\theta} p(w|\theta_{t_i})}$$

$$\begin{aligned} &z^{(n)}_{d_i,w}(\theta_B) = \\ &\frac{\lambda_B p(w|\theta_B)}{(1-\lambda_B)(\sum\limits_{j' \in r(i)} \pi^{(n-1)}_{i,j'} p(w|\theta_{d_{j'}}) + \pi^{(n-1)}_{i,\theta} p(w|\theta_{t_i})) + \lambda_B p(w|\theta_B)} \end{aligned}$$

In the M-step, given the expectation of the hidden variables, we get the best parameters $p(w|\theta_t)$ as:

$$\begin{aligned} &p(w|\theta_t) \\ &= \frac{\sum\limits_{d_i, t_i = t} c(d_i, w))(1 - z^{(n)}_{d_i,w}(\theta_B)) z^{(n)}_{d_i,w}(\theta_t) + \mu_E p(w|\theta_0)}{\sum\limits_{w'} \sum\limits_{d_i, t_i = t} c(d_i, w'))(1 - z^{(n)}_{d_i,w}(\theta_B)) z^{(n)}_{d_i,w'}(\theta_t) + \mu_E p(w'|\theta_0)} \end{aligned}$$

By integrating Lagrange multipliers [29] f_i for each $d_i \in \mathcal{D}$, the inference of $\pi(i)$ boils down to solve a group of cubic equations:

$$\pi^2_{i,*} + \beta_{i,*}\pi_{i,*} + \gamma_{i,*} = 0, \quad * \in r(i) \cup \{\theta\} \qquad (5)$$

where

$$\begin{aligned} \beta_{i,*} &= \frac{\sum\limits_{*' \neq *} (\mathcal{Q}_{\pi(i)}(*, *') + \mathcal{Q}_{\pi(i)}(*', *))(\pi^{(n-1)}_{i,*'} - \mu_{i,*'})}{2Q_{\pi(i)}(*, *)} \\ &- \mu_{i,*} + \frac{f_i}{\mu_G \mathcal{Q}(i)_{*,*}} \\ \gamma_{i,*} &= -\frac{\sum\limits_{w} c(d_i, w)(1 - z^{(n)}_{d_i,w}(\theta_B)) z^{(n)}_{d_i,w}(\theta_{d_j})}{\mu_G \mathcal{Q}(i)_{*,*}} \end{aligned}$$

Let $\pi_{i,*} = \frac{-\beta_{i,*} + \sqrt{\beta^2_{i,*} - 4\gamma_{i,*}}}{2}$ be the root of Equation 5. It is easy to prove that $\pi_{i,*}$ can be arbitrarily close to zero when $f_i \to +\infty$ and arbitrarily large when $f_i \to -\infty$. Also, the derivative $\frac{\partial \pi_{i,*}}{\partial f_i} = \frac{1}{2}\left(-1 + \frac{\beta_{i,*}}{\sqrt{\beta^2_{i,*} - 4\gamma_{i,*}}}\right) < 0$. Hence, it is guaranteed that there exist valid solutions for the group of equations that satisfy the constraint $\sum_{* \in r(i) \cup \{\theta\}} \pi_{i,*} = 1$ for each d_i in $\mathcal{D}$.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of our TIDE model on synthetic datasets as well as data collected from two real-world social communities, *i.e.*, *DBLP* [43] and *Twitter* [25].

4.1 Experimental Setup

4.1.1 Data Collections

The DBLP Dataset ([43]). The Digital Bibliography and Library Project (DBLP) is a web accessible database of the bibliographic information of computer science publications. In this experiment, we use a collection of DBLP articles augmented with citation information, released by the ArnetMiner group [2], which contains $1,632,442$ publications by $1,741,170$ researchers with $2,327,450$ citations. After filtering out papers without text or citation information, $243,425$ papers and $246,839$ authors are retained. This dataset represents a typical academic community, with a social network of authors (with coauthoring and citation relations) and a collection of scientific papers.

The Twitter Dataset ([25]). Twitter is a well known social networking and microblogging community. In this experiment, the Twitter dataset was crawled down by the DAIS group at University of Illinois , which contains $5,000$ socially connected users and their most recent 200 tweets posted before Nov. 23, 2010. Totally, there are $103,968$ oneway following relations, and $51,032$ pairs of friends (mutual following relations). This dataset represents a typical social community with a directed social network (defined by following relations) and a collection of tweets.

[2] http://arnetminer.org/DBLP_Citation

Synthetic Dataset. The lack of ground truth on real world dataset makes it hard to evaluate the model performance quantitatively. To achieve quantitative evaluation, we construct a synthetic dataset which simulates the diffusion of $1,000$ themes. For each theme, we extract a subgraph of $1,000$ authors from the *DBLP* dataset using breath first search from a random seed author. This subgraph is used to simulate the social network in which the theme diffuses. We then randomly attach $1,859$ empty and time-stamped documents to the authors in this network [3]. We then simulate a diffusion graph of the $1,859$ documents that is regularized by the simulated social network structure. Specially, we first randomly generate a network of the $1,859$ documents using Erdos/Renyi model, with the average degree of 5 (consistent with the real statistics in the DBLP dataset). The direction of each edge is determined by the time stamps of the documents (always points to a "newer" document). We then weight each edge based on the social connections of the authors of the two document plus a random effect. This directed and reweighed random network simulates the real diffusion network among documents. For each theme, we also simulate a sequence of 10 evolving topic snapshots based on the dynamic topic models [2]. Finally, the text content of each document is generated by a simple mixture model with all documents that have "influenced" this document as well as the corresponding topic snapshot.

4.1.2 Baselines

The NetInf Model [11]. NetInf is a typical model that infers the diffusion network of explicit user behaviors. Given the time stamps at which individuals adopt a behavior, *NetInf* identifies the optimal general network of users that best explains the observed adoptions. Comparing to TIDE, *NetInf* is trying to infer the general social network structure according to the observation of the propagations of a group of events, while *TIDE* infers the theme-specific diffusion graph with the help of a general social network. Note that *NetInf* doesn't consider text information, thus cannot track topic evolution.

If we treat each term with a positive probability in the primitive topic as an explicit event/behavior, then a document adopts that behavior explicitly if the term appears in the document. We are then able to infer the optimal document network using *NetInf*. This optimal network is easily converted into a diffusion graph by endowing each edge with equal flow volume.

The IndCas Model [34]. The second baseline is a deviation of the independent cascade model stated in [34], where the probability for an active document to infect another is proportional to the strength of the social connection between their authors with an exponential decay effect [7] (see Section 3.3). We convert these probabilities into a diffusion graph where the diffusion flow from d_j to d_i is proportional to the probability that d_j infects d_i.

The TIDE- Model. To evaluate the effectiveness of social connections in our models, we implement a special version of *TIDE* by removing the regularization term with the network structure, *i.e.*, by setting $\mu_G = 0$.

We believe *NetInf* and *IndCas* are good representatives of diffusion inference models of explicit behaviors, which do not consider textual information or the evolution of topics. *TIDE-* on the other hand ignores the effects of social connections.

4.2 Experiments on Synthetic Data

The goal of the experiments on synthetic data is to quantitatively evaluate how well each method can (i) infer diffusion graphs, (ii) estimate contribution of novelty (if possible), and (iii) discover snapshots in topic evolution (if possible). Given the simulated social community (the social network, the document collection, and the primitive topic), our goal is to recover the diffusion graph and the topic snapshots. The parameters in the *TIDE* model are set empirically as $\mu_E = 10$, $\alpha = 30$, and $\mu_G = 10$.

4.2.1 Analysis on Information Diffusion

We first look at how successful models are at inferring diffusion graphs. Let us first introduce the evaluation metrics.

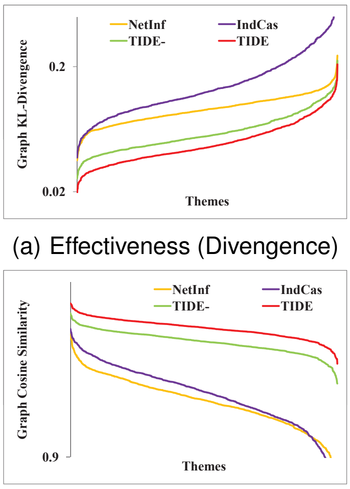
Definition 4.1. **Graph KL-Divergence**. The symmetrized Kullback-Leibler divergence [19] is a classic measure of the difference between two probability distributions. We extend the SKLD and define an evaluation metric to measure the discrepancy between two diffusion graphs $\Pi_{\mathcal{P}}$ and $\Pi_{\mathcal{Q}}$ on the same document collection $\mathcal{D}$:

$$GD_{KL}(\Pi_{\mathcal{P}},\Pi_{\mathcal{Q}}) = \frac{\sum_{d_i\in\mathcal{D}} (D_{KL}(\pi_{\mathcal{P}}(i)||\pi_{\mathcal{Q}}(i)) + D_{KL}(\pi_{\mathcal{Q}}(i)||\pi_{\mathcal{P}}(i)))}{2|\mathcal{D}|}$$

Definition 4.2. **Graph Cosine Similarity**. We also define a metric of similarity between two diffusion graphs $\Pi_{\mathcal{P}}$ and $\Pi_{\mathcal{Q}}$, as the average cosine similarity [41] between their diffusion vectors.

$$Cos(\Pi_{\mathcal{P}},\Pi_{\mathcal{Q}}) = \frac{1}{|\mathcal{D}|}\sum_{d_i\in\mathcal{D}}\frac{\pi_{\mathcal{P}}(i)\cdot\pi_{\mathcal{Q}}(i)}{||\pi_{\mathcal{P}}(i)||\cdot||\pi_{\mathcal{Q}}(i)||}$$

A better model should infer a diffusion graph that is closer to the "ground truth" (the simulated diffusion network), that is, a lower KL-divergence and a higher Cosine similarity.



(a) Effectiveness (Divengence)

(b) Effectiveness (Similarity)

Figure 2: Diffusion Evaluation on the Synthetic Dataset

In practice, we calculate the two metrics for the result of each method and each theme[4], and connect the KL-divergence scores in decreasing order (Figure 2(a)) and Cosine-similarity scores in increasing order (Figure 2(b)). The aggreated performance of the $1,000$ themes is reported in the 1^{st} and 2^{nd} columns of Table 1.

[3] According to the statistics on our *DBLP* datset, each researcher has 1.859 first-authored publications in average.

[4] To make the results from all methods comparable, vertices associated with topic snapshots are removed from the diffusion graphs inferred by *TIDE* and *TIDE-*.

We can conclude that *TIDE* achieves the best performance, then *TIDE-*, then *NetInf*, and then *IndCas*.

4.2.2 *Proof of Combined Power*

With this experiment, we can also prove that both social networks and text information play an important role the inference of topic diffusion.

Object	TDG		$\mathcal{H}$		$\mathcal{G}$	
Metric	GKLD	GCS	GKLD	GCS	GKLD	GCS
NetInf	0.0936	0.9359	1.2906	0.8685	0.9490	0.8022
IndCas	0.1601	0.9313	1.2971	0.8550	**0.7210**	**0.8975**
TIDE–	0.0628*	0.9691*	**0.9906**	**0.9494**	0.8757	0.8407
TIDE	**0.0524***	**0.9722***	1.0109	0.9378	0.8459	0.8547

Table 1: Diffusion Evaluation on the Synthetic Dataset (TDG = True Diffusion Graph, GKLD = Graph Kullback-Leibler Divergence, GCS = Graph Cosine Similarity)

We measure the statistical significance of the improvement using the dependent t-test. * means that the improvement (over the row above) hypothesis is accepted at significance level 0.001.

First, we create a document network (denoted as $\mathcal{H}$), where the edge weight is proportional to the content similarities between documents. We compare each inferred diffusion graph with $\mathcal{H}$, and report the aggregated value of the two metrics in the 3^{rd} and 4^{th} columns of Table 1.

Second, we project each diffusion graph Π into a user network (denoted as $f(\Pi)$), compare $f(\Pi)$ with the general social network $\mathcal{G}$, and report the aggregated value of the two metrics in the last two columns of Table 1.

We can observe some phenomena that accord with our hypothesis in designing our model: *TIDE-* infers diffusion graphs only considering textual information without considering the social network structure, while *IndCas* infers the diffusion network purely based on the social influences. Indeed, the diffusion networks inferred by *TIDE-* are significantly biased towards the document similarity networks $\mathcal{H}$, and the diffusion networks inferred by *IndCas* are biased towards the social networks $\mathcal{G}$. **Neither of them infers diffusion networks that are closer to the ground truth than TIDE, which employs both text information and the social network**.

4.2.3 *Analysis on Content Evolution*

In this experiment, we study how successfully TIDE and the baseline models track topic evolution. Since *NetInf* and *IndCas* are not able to handle topics, we compare our models *TIDE* and *TIDE-* with a simple mixture model stated in [48].

We repeat similar experiments as done in Section 4.2.1. We use two similar metrics (*i.e.*, the symmetrized KL-divergence and the Cosine similarity) to measure the closeness of the discovered word distributions of the topic snapshots to the "ground truth" (topic snapshots we construct in the synthetic dataset). The results are reported in Table 4.

As shown above, *TIDE* outperforms the other two methods with sufficient certainty, which proves our statement in Section 1: **the evolution and the diffusion of topics are compound processes; the success of one aspect will help the inference of the other**.

Metric	KLD	CS
FM	0.4281	0.7033
TIDE-	0.3301*	0.8622*
TIDE	**0.2893***	**0.8774***

Table 4: Evolution Evaluation on the Synthetic Dataset (KLD = Kullback-Leibler Divergance, CS = Cosine Similarity, FM = Feedback Model [48])

* means the improvement (over the above row) hypothesis is accepted at the significance level 0.001 based on dependent t-test.

4.3 Experiments on Real Social Networks

We present the experiments on real world social communities in this section. Note that "ground truth" diffusion networks and topic snapshots are usually not available.

4.3.1 *Verifying Motivating Observations*

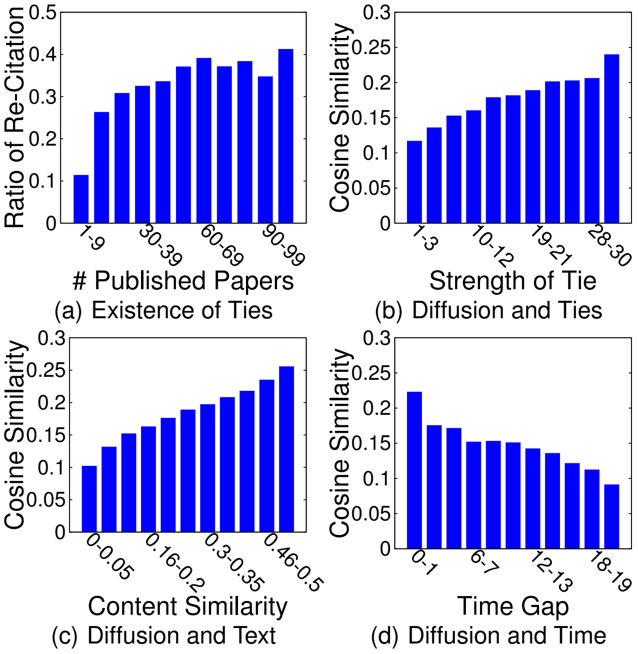


Figure 3: Verifying Observations by DBLP-Citation Dataset

We start with the verification of the authenticity of the three motivating observations stated in Section 3.1. We expect that social influence, so that an author is more likely to adopt topics from the documents of her social connections. If this is the case, an author will pay consistent attention to papers published by authors she knows, or she has cited before. One intuitive way to verify this is through the behavior of 're-citation', *i.e.*, once the author cited one paper, it is likely that she will cite the paper of the same author again. We group authors by the number of publications, and plot the average ratio of re-citation in Figure 3(a). It shows that there are substantial re-citation behaviors, when an author publish more papers, the ratio of re-citation also grows. This verifies the existence of social influence in document-level information diffusion.

Instead of inferring the diffusion, a rough proxy of the influence of a cited paper d_r on a citing paper d_c can be measured by the author's behaviors after the citation. Generally, if the authors of d_c

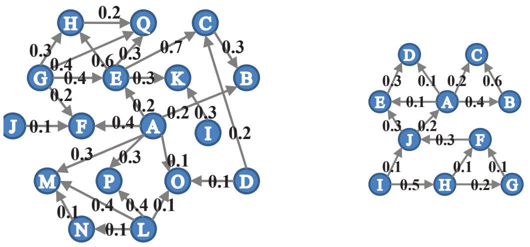
ID	Publication	ID	Publication
A	J. Han, *etc*, SIGMOD'00.	B	A. Khan, *etc*, KDD'10.
C	X. Yan, *etc*, SIGMOD'04.	D	M. Zaki, *etc* KDD'03.
E	X. Yan, *etc*, KDD'03.	F	Y. Chi, *etc*, TKDE'05.
G	M. Zaki, KDD'02.	H	A. Bifet, *etc*, KDD'08.
I	X. Yan, *etc*, KDD'05.	J	U. Rükert, *etc*, SAC'04.
K	C. Chen, *etc*, CIKM'08.	L	J. Wang, *etc*, KDD'03.
M	J. Wang, *etc*, TKDE'05.	N	F. Pan, *etc*, KDD'03
O	A. Lee, *etc*, Infomation System'10.		
P	U. Yun, Knowledge-Based System'08		
Q	J. Balc*á*zar, *etc*, Machine Learning'10.		

Table 2: Publications Shown in Figure 4(a)

ID	Tweet (incomplete)
A	Inception had better special effects than Videodrome.
B	Inception's effects might take some Oscars.
C	I predict Inception's 12 Oscar nominations.
D	It has to be like a 3rd level Inception dream.
E	I wonder what level of recursive dreams.
F	Inception. What a brilliant, mind-twisting movie.
G	Watching inception. Long movie.
H	You'd be odd on twitter if you haven't seen Inception.
I	First time I have seen a movie in a theater in the last 6 months.
J	If you like intelligent movies and complex plots, go to see Inception.

Table 3: Tweets Shown in Figure 4(b)

publish many papers related to d_r after they publish d_c, it is fair to believe d_r is quite influential to d_c. We partition the citations (each of which is recognized by a cited paper d_r and a citing paper d_c) into different groups according to the strength of the social connection between their authors. For each citation, we then compute the average document similarity between d_r and all papers published by d_c's authors after they had published d_c. The aggregated similarity is plotted in Figure 3(b). We repeat the same experiment, but partitions citations by degrees of the content similarity of d_r and d_c (Figure 3(c)), as well as the time gap (3(d)) between d_r and d_c. Figure 3(b)-3(d) prove our motivations that the (proxy) influence between two documents increases with the strength of social ties (Observation 2) and the content similarity, but decays over time (Observation 3).



(a) Theme 1 (DBLP-Citation) (b) Theme 2 (Twitter)

Figure 4: Case Study on Real Networks: Diffusion Graphs

4.3.2 Case Study

We select two themes for case study: one is about the research topic *'frequent pattern mining'* on the DBLP-Citation dataset (see the 1^{st} column of Table 5), and the other is about the movie *'inception'* on the Twitter dataset (see the 1^{st} column of Table 7).

Analysis on Information Diffusion. For theme 1, we apply the *TIDE* model on 344 papers published during the past ten years (2000 to 2010), which contain at least three primitive keywords in the title or abstract. A subgraph of the diffusion graph estimated by TIDE is shown in Figure 4(a), on a subset of 17 selected papers (listed in Table 2). The volume of each diffusion flow is marked on the edge. To quantitatively access the result, we compare the graph with three alternative "diffusion graphs." In the first graph, the weight of an edge $d_r \rightarrow d_c$ is set proportional to the total length of citation sentences where d_c mentions d_r. We then employ two experts to manually score the impact of each reference paper in a scale from one to five. The **Mean Absolute Error** [35], as the statistical metric of accuracy, based on each criteria, and the **Cohen's Kappa Coefficiency** [37], as the measure of inter-criteria agreement, are reported in Table 10.

Theme 2 has been used as the running example in Section 1 (see Figure 1), and let us reveal more details. We apply the *TIDE* model on 361 tweets containing the keyword *'inception'*, and draw the diffusion graph on 10 selected tweets (listed in Table 3) in Figure 4(b). We repeat the same evaluation procedure as done for theme 1 (see Table 11), only except that the edge weight of the first criteria graph is decided by whether one tweet was replying the other.

MAE	SL	Exp1	Exp2
TIDE	0.1217	**0.1080**	0.1195

CKC	Exp1	Exp2
SL	0.5019	0.2095
Exp1	–	**0.6333**

Table 10: Theme 1 (DBLP-Citation)

MAE	RR	Exp1	Exp2
TIDE	0.3632	**0.1301**	0.1351

CKC	Exp1	Exp2
RR	0.3583	0.3726
Exp1	–	**0.7500**

Table 11: Theme 2 (Twitter)

Table 12: Case Study on Real Networks: Accuracy Evaluation (MAE = Mean Absolute Error, CKC = Cohen's Kappa Coefficiency, SL = Sentence Length, RR = Replying Relation)

In both cases, the opinions of the first expert gains the most agreement from others, and our result has the highest accuracy against the truth ground supplied by the first expert.

Analysis on Content Evolution. We apply both *TIDE* and the feedback model [48] to extract the topic snapshots for two themes. Top words (with probabilities) of several selected topics are listed in Table 5-8. Note, to demonstrate more results, the word *'fre-*

Primitive Topic		Year 2003		Year 2005		Year 2009	
frequent	0.20	itemset	0.05	itemset	0.04	itemset	0.03
pattern	0.40	GSM	0.03	tree	0.02	tree	0.02
mining	0.20	association	0.02	parallel	0.01	sequence	0.01
graph	0.05	apriori	0.02	graph	0.01	graph	0.01
tree	0.05	tree	0.01	sequence	0.01	slide	0.01
sequence	0.05	graph	0.01	traversal	0.01	gram	0.01
itemset	0.05	subgroup	0.01	optimize	0.01	window	0.01
		sequential	0.01	suffix	0.01	apriori	0.01

Table 5: Topic Snapshots by *TIDE* on Theme 1 (DBLP-Citation)

Year 2003		Year 2005		Year 2009	
efficient	0.02	close	0.01	sequential	0.02
close	0.01	itemset	0.01	itemset	0.01
association	0.01	match	0.01	tree	0.01
support	0.01	tree	0.01	graph	0.01
query	0.01	graph	0.01	database	0.01
temporal	0.01	sequential	0.01	efficient	0.01
graph	0.01	efficient	0.01	rule	0.01
rule	0.01	application	0.01	match	0.01

Table 6: Topic Snapshots by [48] on Theme 1

Primitive Topic		Jul 16-19		Jul 20-23		Jul 24-27	
inception	1.00	watch	0.05	dream	0.06	oscar	0.04
		night	0.05	mind	0.05	effect	0.04
		movie	0.05	level	0.03	dream	0.02
		special	0.03	walk	0.01	clever	0.01
		enjoy	0.01	recursive	0.01	briliant	0.01

Table 7: Topic Snapshots by *TIDE* on Theme 2 (Twitter)

Jul 16-19		Jul 20-23		Jul 24-27	
movie	0.06	type	0.05	oscar	0.03
night	0.06	eye	0.05	act	0.03
special	0.03	watch	0.05	dream	0.03
watch	0.03	night	0.05	strong	0.02
bad	0.03	dream	0.04	night	0.02

Table 8: Topic Snapshots by [48] on Theme 2

Table 9: Case Study on Real Networks: Topic Evolution

quent', *'pattern'* and *'mining'* are eliminated from Table 5 and 6; and the word *'inception'* are eliminated from Table 7 and 8.

As elaborated in Section 3.2, the topic component of *TIDE* utilize (i) a background model to absorb common words, and (ii) reference models to absorb old words, so that topic snapshots would attract more discriminative and meaningful words that describe the novel aspect of a theme. For example, the topic at *'Year 2009'* in Table 6 reveals a new trend of mining patterns up to certain length (*i.e.'gram'*) in a *'sliding' 'window'*, and the topic at *'Jul 20-23'* in Table 8 talks about the movie plots such as the *'level'* of a *'recursive' 'dream'*. However, since [48] only considers the idea of background model, these interesting new words are easily overlooked, because antiquated words, such as *'efficient'* in Table 5 and *'watch'* in Table 7, repeatly appear in lots of topic snapshots.

5. RELATED WORK

TIDE is a novel probabilistic model for the joint inference of diffusion and evolution of topics, by comprehensively considering 1) the generation of text, 2) the effects of social networks, 3) contribution of novelty, and 4) the topic evolution. To the best of our knowledge, there is no existing model that considers all these factors in a unified and principled way. There are, however, several lines of related work.

Information Diffusion. Information diffusion [1, 13, 18, 44, 24, 9, 21, 22] is a classic topic in social network analysis, which models the cascade of behaviors on a network structure. [6, 12, 20, 17] aim at a subset of nodes or links in a network that could maximize (or minimize) the spread. [16, 42, 45, 27] estimate social graphs with edges labeled with probabilities of influence between users. [11] infers a latent network structure, over which events spread well. [18] predicts the sharing scale of an opinion. This line of work, however, usually do not consider textual topics evolving along time, and thus hard to be applied to our tasks.

Topic Modeling. Topic modeling approaches [14, 3] have been developed to mine variations of topics [39, 23, 8, 31, 10]. Specially, incorporating network regularization into topic modeling has become a popular tendency, such as NetPLSA [30], Laplacian PLSI [4], iTopicModel [40] and locally-consistent Models [5, 26]. [30] uses a harmonic function to enforce the constraint that topic distribution on neighboring nodes should be similar, [40] defines a Markov Random Field on the graph to model the influence between nodes in a generative way, and [25] leverages Gibbs Random Field to estimate the popularity index of a textual topic depending on social connections and history. [45, 42] model the influence graph among users by considering both network structures and topics. However, none of these methods aim to infer diffusion paths among documents. Thus they can not be directly applied to our problem.

6. CONCLUSION

In this paper, we propose *TIDE*, a novel probabilistic model for the joint inference of diffusion and evolution of topics in social communities. TIDE integrates the generation of text, the evolution of topics, and the social network structure in a unified model. Given the primitive form of any arbitrary topic, *TIDE* effectively tracks the topic snapshots that evolves along time and reveals the latent diffusion paths of the topic. Comprehensive experiment studies on both synthetic data and two real-world datasets show that TIDE outperforms existing approaches.

One important finding is that the discovery of topic diffusion and topic evolution benefits significantly from the joint inference process. Social influence still plays an important role in the diffusion of topics. Both text information and the general social network structure play an irreplaceable role to the inference process. We expect a future extension of TIDE that model the evolution of the social network structure in addition to the evolution of topics.

7. REFERENCES

[1] L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54, 2006.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *NIPS*, pages 601–608, 2001.

[4] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[5] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, page 14, 2009.

[6] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038, 2010.

[7] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *the National Academy of Sciences*, 105(41):15649-15653, Oct 2008.

[8] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML*, pages 233–240, 2007.

[9] D. Easley and J. Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. *Cambridge University Press*, 2010.

[10] S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *ICML*, pages 375–382, 2010.

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

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

[13] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[14] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[15] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[16] M. Kimura, K. Saito, and H. Motoda. Minimizing the spread of contamination by blocking links in a network. In *AAAI*, pages 1175–1180, 2008.

[17] M. Kimura, K. Saito, and R. Nakano. Extracting influential nodes for information diffusion on a social network. In *AAAI*, pages 1371–1376, 2007.

[18] M. Kimura, K. Saito, K. Ohara, and H. Motoda. Learning to predict opinion share in social networks. In *AAAI*, 2010.

[19] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[20] C. Lee, H. Kwak, H. Park, and S. B. Moon. Finding influentials based on the temporal order of information adoption in twitter. In *WWW*, pages 1137–1138, 2010.

[21] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *TWEB*, 1(1), 2007.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

[23] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, pages 577–584, 2006.

[24] D. Liben-Nowell and J. Kleinberg. Tracing the flow of information on a global scale using internet chain-letter data. *the National Academy of Sciences*, 105(12):4633–4638, 2008.

[25] C. X. Lin, B. Zhao, Q. Mei, and J. Han. Pet: a statistical model for popular events tracking in social communities. In *KDD*, pages 929–938, 2010.

[26] J. Liu, D. Cai, and X. He. Gaussian mixture model with local consistency. In *AAAI*, 2010.

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

[28] B. R. M. M. H. MacRoberts. Problems of citation analysis. *Scientometrics*, 36(3):435–444, 1996.

[29] G. McLachlan and T. Krishnan. The em algorithm and extensions. *Wiley series in probability and statistics, Hoboken*, 2008.

[30] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, pages 101–110, 2008.

[31] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207, 2005.

[32] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *WWW*, pages 727–736, 2006.

[33] H. Rue and L. Held. Gaussian markov random fields: Theory and applications - theory and application. *Chapman and Hall/CRC*, 2006.

[34] K. Saito, M. Kimura, K. Ohara, and H. Motoda. Selecting information diffusion models over social networks for behavioral analysis. In *ECML/PKDD (3)*, pages 180–195, 2010.

[35] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

[36] A. Si, H. V. Leong, and R. W. H. Lau. Check: a document plagiarism detection system. In *SAC*, pages 70–77, 1997.

[37] N. Smeeton. Early history of the kappa statistic. *Biometrics*, 41:795, 1985.

[38] H. W. Sorenson. Parameter estimation: Principles and problems. *Marcel Dekker*, 1980.

[39] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, pages 306–315, 2004.

[40] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *ICDM*, pages 493–502, 2009.

[41] P. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. volume ISBN:0321321367, 2005.

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

[43] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.

[44] X. Wan and J. Yang. Learning information diffusion process on the web. In *WWW*, pages 1173–1174, 2007.

[45] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[46] J. Whittaker. Graphical models in applied multivariate statistics. In *Wiley Publishing*, 2009.

[47] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.

[48] C. Zhai and J. D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410, 2001.