Social Network Analysis and Mining (2021) 11:1
https://doi.org/10.1007/s13278-020-00705-z

ORIGINAL ARTICLE

Probabilistic reasoning system for social influence analysis in online social networks

Lea Vega[1] · Andres Mendez-Vazquez[1] · Armando López-Cuevas[2]

Received: 16 December 2019 / Revised: 26 August 2020 / Accepted: 28 October 2020 / Published online: 19 November 2020

Abstract

People interact with other people in their daily life, either for work or for personal reasons. These interactions are often complex. Thus, interactions that an individual has with other individuals, to some extent, influence the decisions they make. There have been many efforts to uncover, explore, and measure the concept of social influence. Thus, modeling influence is an open and challenging problem where most evaluation models focus on online social networks. However, they fail to characterize some properties of social influence. To address the limitations of the previous approaches, we propose a novel Probabilistic Reasoning system for social INfluence analysis (PRIN) to examine the social influence process and elucidate the factors that affect it in an attempt to explain this phenomenon. In this paper, we present a model that quantitatively measures social influence in online social networks. Experiments on a real social network such as Twitter demonstrate that the proposed model significantly outperforms traditional feature engineering-based approaches. This suggests the effectiveness of this novel model when modeling and predicting social influence.

Keywords Social influence analysis · Social networks · Bayesian network

1 Introduction

Social influence (SI) is a common feature of everyday life: we either try to influence others or are influenced by them many times each day. For example, colleagues have a strong influence on one's work, while friends have a strong influence on one's daily life. This influence can be somewhat banal such as what type of shoes to buy—or more significant—such as whether to vote for one candidate versus others. Rashotte (2007) defined SI as a change in an individual's thoughts, feelings, attitudes, and behaviors that results from interaction with another individual or a group. Influence is an invisible, complex, and subtle phenomenon that governs social dynamics and user behaviors. Besides, SI takes many different forms and can be seen in processes of conformity, socialization, peer pressure, obedience, leadership, persuasion, minority influence, among others (Han and Li 2018).

A social network (SN) is a social structure made up of individuals or organizations, which are connected by one or more specific types of interdependency, such as friendship, kinship, common interest, likes/dislikes, or relationships of beliefs, knowledge or prestige (Travers and Milgram 1967). As the Internet evolved, online social networks (OSNs) emerged such as Twitter, LinkedIn, and Facebook. They have attracted a lot of attention since they allow users to share ideas, activities, events, and interests within their networks. In contrast to traditional (offline) social networks, OSNs store a register of the interaction between users based on their content shared between them, and how this content is propagated on the Internet, which is a result of SI.

Many researchers have tried to test or examine whether there is an influence, and how people influence each other in SN. Understanding how users influence each other and quantifying it can benefit various applications. In the field of data mining, some recognized applications include viral marketing (or targeted advertising in general) (Freeman 1978; Brown and Reingen 1987), recommender systems

✉ Lea Vega
lea.vegar@cinvestav.mx

Andres Mendez-Vazquez
andres.mendez@cinvestav.mx

Armando López-Cuevas
acuevas@itesm.mx

[1] Research Center for Advanced Study from the National Polytechnic Institute, Zapopan, Mexico

[2] Technological and Higher Education Institute of Monterrey, Monterrey, Mexico

(Pálovics et al. 2014), analysis of information diffusion in social media like Twitter and Facebook (Tang et al. 2009), events detection (Weng et al. 2010), experts finding (Franks et al. 2013), link prediction (Qiu et al. 2018), and ranking of feeds.

Other disciplines outside social psychology, such as computation, politics, marketing recognize that social influence is key to their concerns. However, they are building models of influence largely without recognizing the extensive conceptual and empirical work done by social psychologists. Typical models in SIA such (Albert and Barabási 2002; Faloutsos et al. 1999; Newman 2003; Strogatz 2001; Tang et al. 2009; Hu et al. 2013, 2015; Cai et al. 2017; Qiu et al. 2018) focus on macro-level models such as degree distributions, diameter, clustering coefficient, and communities or fail to capture the complexity in the SI process. However, in online real situations, the spread of influence occurs through populations over a span of time with each individual serving as both a source and a target showing special properties.

Therefore, novel methods are required to characterize and quantify the process of the SI that can be extended to OSNs, which incorporate social theories to create a model that portrays this phenomenon as close to reality. Thus, the general objective of this research is to design and implement a novel model that characterizes the social influence process based on sociological theories and probabilistic reasoning theory for quantifying each user over the other users inside an online social network. The following specific objectives are defined for this work: First, to analyze the social influence process from the social science point of view in online social networks for selecting a set of involved factors; Second, to formalize the concept of social influence and sociological theories about social influence, such as user's behavior, user's profile, user's temporal variation interest, and temporal evolution of relationships; Third, to design a mathematical model that integrates the formalized factors to represent the social influence process; Finally, to generate a probabilistic reasoning system with the designed model and an existing inference algorithm to ask queries to the model.

In the purse of these objectives, we present a novel algorithm called PRIN, a generative model describing the dynamic of the social influence process. Thus, PRIN allows measuring topical user-level social influence strength through the modeling of the previous objectives. Our final aim is to bring together large volumes of unstructured data such as content shared, and heterogeneous information such as structural link and diffusion links in this model. Additionally, research was performed in social sciences such as physiology, sociology, and anthropology. Therefore, PRIN reflects a broader and deeper knowledge about human behavior in its mathematical inception. We conducted an extensive set of experiments to evaluate the effectiveness of our model to discover user influence. The final results suggest that our model can improve the prediction performance, thus allowing to mine additional information inside social networks.

The rest of this paper is organized as follows: Sect. 2 summarizes related work about social influence analysis. Section 3 formulates social influence problem. Section 4 introduces the proposed model in detail. Section 5 presents the proposed social influence analysis architecture. Sections 6 and 7 present the experimental results and discussion of the validation of the model. Finally, Sect. 8 presents the conclusion and future work of this work.

2 Related work

Social influence analysis has been widely studied in the literature. We have explored related work in two main dimensions (March 1955). The first dimension is the explanatory models, and the second is predictive models for social influence in online social networks. We briefly discuss them, respectively, in this section.

2.1 Explanatory models

The explanatory models aim to examine the social influence process and elucidate the factors that affect it in an attempt to explain this phenomenon. Yang and Leskovec (2010) addressed influence as a form of information diffusion with temporal dynamics. Wen and Lin (2010) show that combining different social media improves the social influence measure. Crandall et al. (2008) studied the correlation between social similarity and influence. However, they focus on qualitative identification of influence existence, and they do not provide a quantitative measure of the influential strength. Xiang et al. (2010) and Goyal et al. (2010) further investigate how to learn the influence probabilities from the history of user actions. However, these methods either do not consider the influence at the topic-level or ignore the influence propagation. Kempe et al. (2003) focused on how influence propagated across a network, assuming an influential user is one whose initial adoption would eventually result in the most number of total adoptions by all users. However, they did not consider the topic-level influence. Hu et al. (2015) modeled topics and communities in a unified latent framework and extract inter-community influence dynamics. Cai et al. (2017) formalized the concepts of community profiling by its internal content profile and external diffusion profile. Besides the global effect of influence, many efforts have been made to estimate the concrete influence strength between individual nodes. Bi et al. (2014) integrate both content topic discovery and social influence analysis based on structural links in the same generative process at the user level.

2.2 Predictive models

The predictive models are used to predict the future social influence process in social networks based on certain factors. Researchers have addressed the problem of measuring social influence by predicting a user's ability to disseminate information, such as retweet behavior on Twitter. Yang et al. (2010) analyzed the effects of different factors on retweeting behavior. Based on their observations, they proposed a semisupervised framework on a factor graph model to predict users' retweet behaviors. Pezzoni et al. (2013) discussed how structural factors and retweet behavior affect information diffusion. Considering the retweeting behavior as atomic behavior, they proposed an agent-based information propagation model to generate a cascade. Fei et al. (2011) were the first to adopt microeconomics methods for social media behavior prediction. Saito et al. (2008) focused on the user-level mechanism in social influence where its near neighbors only influence each user. Tang et al. (2009) proposed a topic-specific influence measure. However, this article assumes topic distribution.

Although previous researches have applied many models to analyze social influence, our work is very different. The main contributions are summarized as follows: (1) another approach and formalization of the social influence in online social networks; (2) a novel probabilistic graphical model that characterizes the social influence process applies to online social networks based on sociological theories and previous attend for modeling this phenomenon; (3) a probabilistic reasoning system for quantifying social influence in probabilistic term; (4) a new architecture for social influence analysis focus on diffusion prediction between users inside social networks.

3 Problem formulation

The problem of social influence modeling has been open for a long time since it was proposed. It is because social influence is a relatively subjective concept, and it lacks a universally acknowledged definition. People are frequently confused by the concepts and the measuring methods of social influence. For different social networks, social influence is modeled quite differently. For example, in an electronic commerce network, the most influential users are the ones who can recommend most customers to purchase-specific products successfully (Li et al. 2017). After analyzing the characteristic of social influence and the typical characteristics of OSNs, we introduce a series of formal definitions about the concept of social influence. Then, we formulate the problem of topical mining influence in OSNs.

3.1 Concept definitions

The notation used in this paper is listed in Table 1.

Definition 1 *Social network.* A social network is 3-tuple $\mathcal{G} = (\mathcal{U}, \mathcal{M}, \mathcal{E})$, where $u \in \mathcal{U}$ is an user and $m \in \mathcal{M}$ is a user published message, and $e \in \mathcal{E}$ is a link. The link set $\mathcal{E}$ denotes interaction between users, and it can be derived from different types of user interactions, which are application dependent.

Definition 2 *Topic.* A topic $k \in \mathcal{K}$ is a $|W|$-dimensional multinomial distribution ω_k over a vocabulary, where each component $\omega_{k,w}$ is the probability of a word $w \in W$ belonging to k.

Definition 3 *Message.* A message $m \in \mathcal{M}_u$ contains a set of words w from a given vocabulary, along with a time stamp $t_{u,i}$, meaning that $m_{u,i}$ is the ith message of the user u.

A link is a directed connection $e_{u,v} \in \mathcal{E}$, representing a communication from user u to v. We define two types of links in $\mathcal{E}$: structural links between users and diffusion links between users' messages.

Definition 4 *Structural link.* Let $s_{u,v} \in \mathcal{S}$ be a structural link from user u to user v defined by a directed edge in a social graph. For each possible edge $s_{u,v} \in \mathcal{S}$, if there exists an edge between u and v, $s_{u,v} = 1$; otherwise $s_{u,v} = 0$. Every user has a unique preference to generate structural links with other users based on the interest that a user u has in the published content or non-content of user v. As a result, each $s_{u,v}$ is associated with:

- a Bernoulli distribution $\mu_{u,v}$, which characterizes the user u preference for following v based on the v shared content or non-content.
- a binary variable $y_{u,s}$. When $y_{u,s} = 1$ indicates that the link creation is based on the user u content, whereas $y_{u,s} = 0$ means that content has nothing to do with the link.

Definition 5 *Diffusion link.* Let $d_{m_{u,i},m_{v,j}} \in \mathcal{D}$ be a diffusion link, where message i of user u is a broadcast of message j of user v. Each d is associated with a Bernoulli distribution $\tau_{m,k,t}(u, v)$ characterizing diffusion probability of message $m_{v,*}$ about topic k at time t from user u. This link represents a change in the state of u due to the influenced exerted by v.

Definition 6 *Profiling vector.* The profiling vector $\mathbf{p_u}$ of user $u \in \mathcal{U}$ is defined as an augmenting vector, where each dimension represents some encoded feature associated with

Table 1 Notation used in the model

Symbol	Description
$\mathcal{U}, \mathcal{M}, \mathcal{T}, \mathcal{C}, \mathcal{K}$	Set of users, messages, links, time slices, communities, and topics
$\mathcal{E}, \mathcal{S}, \mathcal{D}$	Set of links, structural links, and diffusion links
$\mathcal{M}_u$	The set of messages published by user u
$m_{u,i}$	The ith message published by user u
$t_{u,i}$	The timestamp of message $m_{u,i}$
$w_{u,i,j}$	The jth word in message $m_{u,i}$
$z_{u,i}$	The topic associated with message $m_{u,i}$
$c_{u,i}$	The community associated with the message $m_{u,i}$
$s_{u,v}$	A structural link from user u to user v
$y_{u,v}$	Binary indicator over the reasons of the structural link $s_{u,v}$ content or non-content related
$d_{m_{u,i},m_{v,j}}$	A diffusion link from message $m_{v,j}$ to message $m_{u,i}$
ω_k	Multinomial distribution over a vocabulary specific to topic k
μ_u	Bernoulli distribution over $\mathcal{S}_u$ specific to user u
χ_i	Multinomial distribution over all the users in the social network
ρ_u, $\rho_{u,k}$	Multinomial distribution over topics specific to user u, represents the interest of u in topic k
π_u, $\pi_{u,c}$	Multinomial distribution over communities specific to user u, membership probability of u in community c
θ_c	Multinomial distribution over topics specific to community c
$\psi_{k,c}$	Multinomial distribution over time specific to topic k and community c
$\alpha, \beta, \gamma, \epsilon, \sigma$	Parameters of the Dirichlet (Beta) priors on Multinomial (Bernoulli) distributions
$\tau_{k,t}(u, v)$	Influence probability (diffusion probability) from user u to user v on topic k at time t

a user, such as a gender, location, topic interest, among others.

Definition 7 *Users similarity*. Let u and v be two users with its associated profiling vectors $\mathbf{p_u}$ and $\mathbf{p_v}$. The user similarity of u and v is defined as the cosine similarity of theirs profiling vectors:

$$\rho(u, v) = \frac{\mathbf{p_u} \cdot \mathbf{p_v}}{\|\mathbf{p_u}\| \|\mathbf{p_v}\|}. \tag{1}$$

Due to homophily, users tend to form communities (Rashotte 2007). Community is a collection of users with similar behaviors and more intense interactions among its members than the rest of the global network (Bi et al. 2014). It can be characterized not only by interaction link structures but also by the content generated by its members.

Definition 8 *Community*. A community $c \in \mathcal{C}$ is associate with:

- a $|\mathcal{K}|$-dimensional multinomial distribution θ_c over topics, where each component $\theta_{c,k}$ represents the probability of c discussing topic k. This topic distribution represents the community different topical interest.
- a $|\mathcal{C}|$-dimensional multinomial distribution $\psi_{k,c}$ over time specific to topic k and community c (Hu et al. 2015).

In social networks, users have multiple roles and are influenced by different community contexts (Xie et al. 2013). Therefore, we employ the mixed-membership approach for a user definition.

Definition 9 *User*. A user $u \in \mathcal{U}$ is associated with:

- A set of messages $\mathcal{M}_u = \{m_{u,1}, \dots, m_{u,|M_u|}\}$ generated by user u.
- A set of structural links $\mathcal{S}_u = \{s_{u,v} \mid v \in \mathcal{U}\}$. Each link $s_{u,v} \in \mathcal{S}_u$ represents a social relationship between user u and user v.
- A set of diffusion links $\mathcal{D}_u = \{d_{m_{u,i},m_{v,j}} \mid t_{u,i} > t_{v,j}$, where $v \in \mathcal{U}$. Each link $d_{m_{u,i},m_{v,j}} \in \mathcal{D}_u$ represents the diffusion of v's message by u.
- A user similarity distribution function $\rho(u, v)$ over user $v \in \mathcal{U}$, which is defined by the user similarity between u and other users.

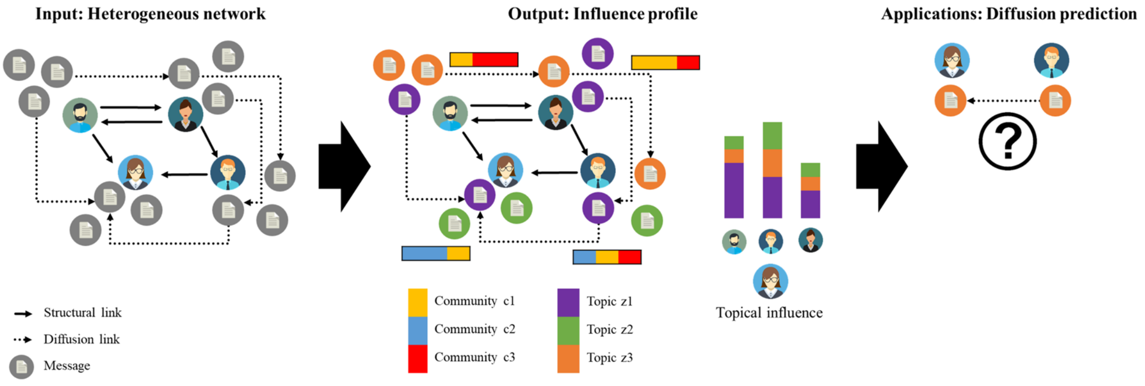


Fig. 1 Overview of our probabilistic reasoning system for social influence analysis

- A $|\mathcal{C}|$-dimensional multinomial distribution π_u over communities, where each component $\pi_{u,c}$ represents the probability of u belonging to community c.
- A $|\mathcal{U}|$-dimensional multinomial distribution over users χ_u, represents the probability of the user u to be considered as important by the other users.

Furthermore, we define user–user influence based on the dynamic process of user behaviors. For example, after user v posts a tweet, one of his followers u reads the tweet and chooses to retweet it. In this case, we assume that user u is influenced by user v, which is reflected as the retweet action performed by u. We model influence in terms of a diffusion probability when a user changes his actual state to another for diffusing specific messages at a particular time.

Definition 10 *User–user topical influence strength*. For a message m about topic k, at time t the user–user topical influence strength is represented as the diffusion probabilities between any two users from u to v, denoted as $\tau_{k,t}(u, v)$.

Please note that the influence is asymmetric, i.e., $\tau(u, v) \neq \tau(v, u)$. Further, we can define the concepts of user influence based on the influence between any pair of nodes.

Definition 11 *User influence*. Let u be a user in a social network; we denote $\tau_t(u)$ as the user influence, which represents the user global influential strength of user u in time t in the social network.

3.2 Problem definition

We have as input a social network, with its observed and latent behavior. As observed variables, there is a set of users $\mathcal{U}$, per user publishes messages $\mathcal{M}$, per user profiling vector ρ, per user's structural links $\mathcal{S}$, and per user's diffusion links $\mathcal{D}$. As a latent behavior, we have factors that according to sociology theories (Rashotte 2007; Granovetter 1977; Liu et al. 2012) constitute the reasons why users are influenced: The per-message topics assignment z, per message community assignment c, per communities' interest θ, per user community membership π, per user preference for building a structural link μ, per user importance in the network χ. These latent factors must be integrated to characterize the social influence process to reason about the social influence process and calculate $\tau_{k,t}(u, v)$ at the end of our observation window.

4 Probabilistic reasoning system for social influence analysis (PRIN)

4.1 Overview of PRIN

Figure 1 shows an overview of the social influence problem. Our goal is to infer topical social influence strength between users. In Fig. 1, we show the input for PRIN: a set of users, each of whom publishes messages; users are connected by structural links and interact with each other by diffusion links, e.g., on Twitter, each user posts tweets, users are connected by followership links, and they retweet each other to diffuse information. Then, for each user, we output a user influence profile as the model's parameters (e.g., user's community membership, user's interest, user's topical influence), and we enable inference about the diffusion probability between users. Finally, we allow applications such as diffusion prediction.

4.2 Probabilistic model design

To model the social influence dynamic, we decided to use a probabilistic Bayesian approach. Bayesian networks are graphical models capable of displaying relationships clearly and intuitively. Further, they handle uncertainty through the established theory of probability. Additionally, they are directional, thus being capable of representing cause-effect relationships. Not only that, they can encode all variables making it possible to handle missing data entries successfully, and they facilitate the use of prior knowledge that we already have about the phenomenon. Besides, the use of Bayesian networks allows us to think of the model in terms of a "generative story" that tells how social influence is created.

Although recent successful attempts inspire some of its building blocks in modeling the dynamics of social networks, including FLDA (Bi et al. 2014), COLD (Hu et al. 2015), and CPD (Cai et al. 2017), PRIN significantly goes beyond those on more comprehensive input features and powerful modeling ability. The social influence process has three main components.

- *User component* Users in a social network usually have multiple community memberships. We associate each user u with a community membership vector π_u. We consider a user u to publish a message $m_{u,i}$ of topic z, due to her community assignment $c_{u,i}$ and the community content profile $\theta_{c_{ui}}$. We assign each post to one community $c_{u,i}$ denoting the community membership of user i when he writes a message.
- *Content component* Each message $m_{u,i}$ shared by a user u contains a bag of words $w_{u,i,1}, \dots, w_{u,i,l}$, where $l =| m_{u,i} |$ denotes the length of the message. In traditional topic models such as Latent Dirichlet Allocation (LDA) (Nallapati et al. 2008), a document is associated with a mixture of topics, and each word has a topic label. This is reasonable for long documents such as academic paper. However, on social media like microblog, a message is usually short and thus is most likely to be about a single topic (Diao et al. 2012). We therefore associate each $m_{u,i}$ to a single latent topic variable $z_{u,i}$ drawn from $\theta_{c_{u,i}}$ to indicate its topic. The words are then generated from the corresponding word distribution $\omega_{z_{u,i}}$.
- *Network component* Every user has a unique preference to form a structural link with others based on content or non-content reason (Bi et al. 2014). For example, A follows B because they are similar, and they both tweet about related topics (content reason), and sometimes A follows C because C is a very famous person (non-content reason). The Bernoulli distribution μ_u characterizes this per user preference. As a result, for the $s_{u,v}$, we first consult μ_u to decide on the value of the binary variable $y_{u,v}$

  - $y_{u,v} = 1$ indicates that the link creation is based on the user's content. In this case, we consider that a user u form a structural link with user v, due to their similarity $\rho_{u,v}$.
  - $y_{u,v} = 0$ indicates that the link creation has nothing to do with the content. In this case, we use χ_v to capture this probability.

Finally, we consider the influence of user v on user u as the probability to form a diffusion link $d_{m_{u,i},m_{v,j}}$ where time for user message $m_{u,i}$ is greater than message $m_{v,j}$, $t_{u,i} > t_{v,j}$. This is drawn from the users' communities membership π, theirs communities' topical interest θ and their structural link probability $s_{u,v}$. The graphical model for PRIN is depicted in Fig. 2, while its generative process is summarized in Algorithm 1.

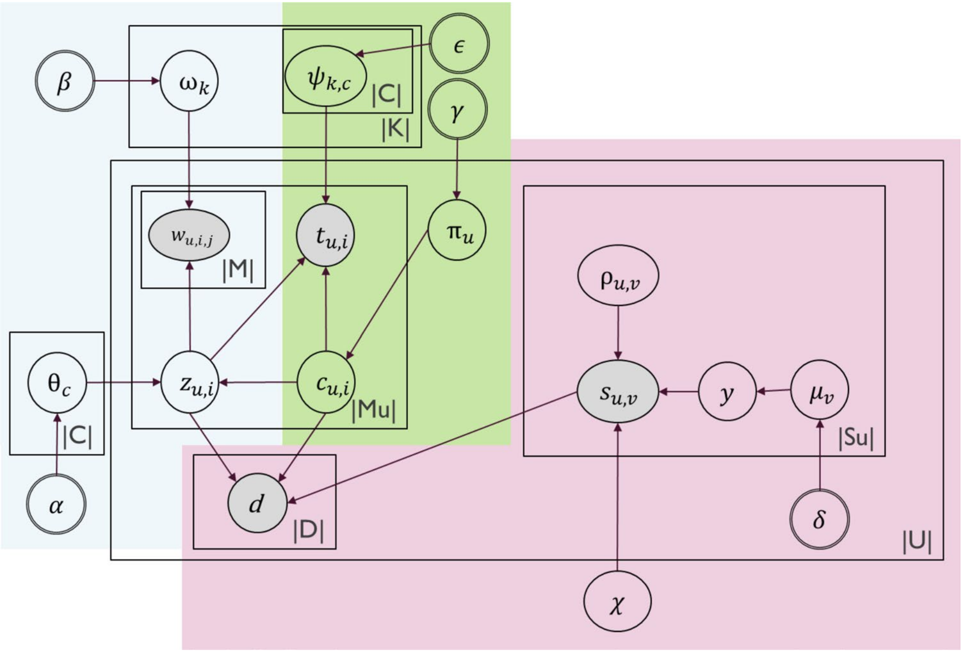


Fig. 2 Plate notation of PRIN model. Network-aware component in pink color, user-aware component in green and content-aware component in blue (Color figure online)

Algorithm 1: Generative process for PRIN

```
Input  : Set of users $\mathcal{U}$, messages $\mathcal{M}$, structural links $\mathcal{S}$, diffusion links $\mathcal{D}$,
         Number of communities $\mid \mathcal{C} \mid$, topics $\mid \mathcal{K} \mid$.
Output: Model parameters.
for each topic $k = 1, \ldots, \mid \mathcal{K} \mid$ do
    Sample its word distribution $\omega_k \mid \beta \sim Dir(\beta)$
    for each community $c = 1, \ldots, \mid \mathcal{C} \mid$ do
        Sample the distribution over time stamps $\psi_{kc} \mid \epsilon \sim Dir(\epsilon)$
    end
end
for each community $c = 1, \ldots, \mid \mathcal{C} \mid$ do
    Sample its topic distribution $\theta_c \mid \alpha \sim Dir(\alpha)$
end
for each user $u = 1, \ldots, \mid \mathcal{U} \mid$ do
    Sample its community membership $\pi_u \mid \gamma \sim Dir(\gamma)$
    Sample its content/non-content preferences $\mu_u \mid \delta \sim Beta(\delta)$
    for each message of user $m_{ui} \in \mathcal{M}_u$ do
        Sample its community assignment $c_{u,i} \mid \pi \sim Mul(\pi_u)$
        Sample its topic assignment $z_{u,i} \mid c_{u,i}, \theta_c \sim Mul(\theta_{c_{u,i}})$
        for each word $w_{u,i,j} \in m_{u,i}$ do
            Sample each word $w_{u,i,j} \mid z_{u,i}, \omega_k \sim Mul(\omega_{z_{u,i}})$
        end
        for each structural link from user $u$ to user $v$ do
            Sample its indicator $y_{u,v} \mid \mu_u \sim Bern(\mu_u)$
            if $y_{u,v} = 0$ then
                Draw its followee distribution $s_{u,v} \mid \chi \sim Mul(\chi)$
            end
            if $y_{u,v} = 1$ then
                Draw its followee distribution $s_{u,v} \mid \rho_{u,v}$
            end
        end
        for each diffusion link of user $u$ from user $v$ do
            Sample $d_{m_{u,i}, m_{v,j}} \mid \theta, \pi_{u,*}, s_{u,v}$
        end
    end
end
```

4.3 Inference task

In the inference task, we aim to infer the latent variables from the observed variables. Therefore, we use collapsed Gibbs sampling for such inference (Guille et al. 2013). Gibbs sampling is an algorithm to approximate the joint distribution of multiple variables by drawing a sequence of samples, which iteratively updates each latent variable under the condition of fixing the remaining variables. Here, we describe the inference algorithm for PRIN based on collapsed Gibbs sampling.

Given a social network $\mathcal{G} = (\mathcal{U}, \mathcal{M}, \mathcal{E})$ and the pre-defined hyper parameters $\alpha, \beta, \gamma, \delta, \epsilon$, PRIN specifies the following full posterior distribution:

$$P(\omega, \psi, \pi, \rho, \theta, \mu, \chi, w, z, c, s, y, d \mid \mathcal{U}, \mathcal{M}, \mathcal{E}, \alpha, \beta, \gamma, \delta, \epsilon). \tag{2}$$

The joint distribution can be decomposed into a product of several factors:

$$\begin{aligned} &P(\omega, \psi, \pi, \rho, \theta, \mu, \chi, w, z, c, s, y, d \mid \mathcal{U}, \mathcal{M}, \mathcal{E}, \alpha, \beta, \gamma, \delta, \epsilon) \\ &\quad \propto \big(P(\omega \mid \beta)P(\pi \mid \gamma)P(\theta \mid \alpha)P(\mu \mid \delta)P(\psi \mid \epsilon) \\ &\qquad P(w \mid \omega, z)P(z \mid \theta, c)P(y \mid \mu)P(c \mid \pi) \end{aligned} \tag{3}$$

$$P(d \mid z, c, s)P(s \mid \rho, \chi, y)P(t \mid psi, c, z)P(\rho), P(\chi)\big). \tag{4}$$

The task of posterior inference for PRIN is to determine the probability distribution of the hidden variables z, c, s, d given the observed words, timestamps, and network. However, exact inference is intractable due to the difficulty of calculating the normalizing constant in the posterior distribution. We use collapsed Gibbs sampling for approximate inference. Because the model uses only conjugate priors (Doucet et al. 2000), we can reduce the number of parameters in the model by integrating out the multinomial distribution $\varphi = \mu, \pi, \theta, \omega, \psi$. This has the effect of taking all possible values of φ into account in our sampler, without

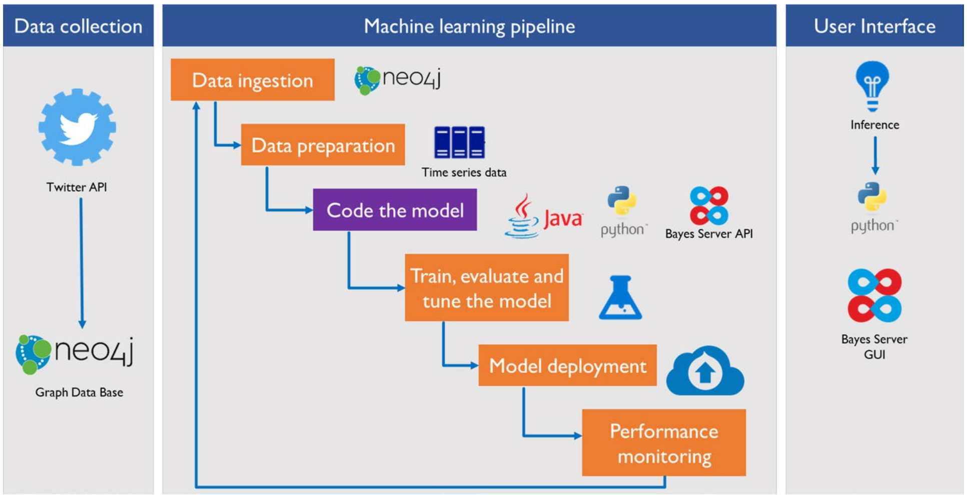


Fig. 3 Architecture of the social influence analysis proposed

representing it as a variable explicitly and having to sample it at every iteration:

$$\begin{aligned}
&P(\omega, \psi, \pi, \rho, \theta, \mu, \chi, d, w, z, c, s, \\
&\quad y \mid \mathcal{U}, \mathcal{M}, \mathcal{S}, \mathcal{D}, \alpha, \beta, \gamma, \epsilon, \sigma) \\
&\quad \propto \int_{\mu} P(y \mid \mu) P(\mu \mid \sigma) \mathrm{d}\mu \\
&\qquad \int_{\omega} P(w \mid \omega, z) P(\omega \mid \beta) \mathrm{d}\omega \\
&\qquad \int_{\pi} P(c \mid \pi) P(\pi \mid \gamma) \mathrm{d}\pi \\
&\qquad \int_{\theta} P(\theta \mid \alpha) P(z \mid \theta, c) \mathrm{d}\theta \\
&\qquad \int_{\psi} P(\psi \mid \epsilon) P(t \mid \psi, c, z) \mathrm{d}\psi \\
&\qquad P(d \mid z, c, s) P(s \mid \rho, \chi, y) P(\rho), P(\chi).
\end{aligned} \tag{5}$$

We are going to focus on $\int_{\pi} P(c_{u,n} \mid \pi_u) P(\pi_u \mid \gamma_u) \mathrm{d}\pi$, this apply for $i \in \varphi$:

$$\begin{aligned}
&\int_{\pi} P(c_{u,n} \mid \pi_u) P(\pi_u \mid \gamma_u) \mathrm{d}\pi \\
&\quad = \int_{\pi} \frac{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u)}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u)} \prod_{i=1}^{\mathcal{C}} \pi_{u,c}^{\gamma_u - 1 + N_i} \mathrm{d}\pi \\
&\quad = \frac{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u)}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u)} \int_{\pi} \prod_{i=1}^{\mathcal{C}} \pi_{u,c}^{\gamma_u - 1 + N_i} \mathrm{d}\pi \\
&\quad = \frac{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u)}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u)} \frac{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u + N_i)}{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_i)}.
\end{aligned} \tag{6}$$

The definition of a Gibbs sampler specifies that in each iteration, we assign a new value to variable $c_{u,j}$ by sampling from the conditional distribution. Intuitively, at the start of an iterations t, we have the collection of all current information at this point in the sampling process. When we want to sample the new value of $c_{u,j}$, we temporarily remove all information about this community from the collection of information $c_{-u,j}$.

$$\begin{aligned}
&P(c_{u,j} = c \mid c_{-u,j}, \gamma_u) = \frac{P(c_{u,j}, c_{-u,j}, \gamma_u)}{P(c_{-u,j}, \gamma_u)} \\
&\quad = \frac{\frac{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u)}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u)} \frac{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u + N_i)}{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_i)}}{\frac{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u)}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u)} \frac{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u + N_{i,-j})}{\Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_{i,-j})}} \\
&\quad = \frac{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u + N_i) \Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_{i,-j})}{\prod_{i=1}^{\mathcal{C}} \Gamma(\gamma_u + N_{i,-j}) \Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_i)} \\
&\quad = \frac{\Gamma(\gamma_u + N^{(c_u,j)}) \Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_{i,-j}^{(c)})}{\Gamma(\gamma_u + N_{i,-j}^{(c_{u,j})}) \Gamma(\sum_{i=1}^{\mathcal{C}} \gamma_u + N_i^{(c)})} \\
&\quad = \frac{\gamma_u + N_{i,-j}^{(c_{u,j})}}{\sum_{i=1}^{\mathcal{C}} \gamma_u + N_{i,-j}},
\end{aligned} \tag{7}$$

where $N_{i,-j}$ is the number message assigned to community i excluding message j.

The same process apply for every element in φ. After sufficient number of sampling iterations. We list the update equations for each variable as below:

$$\begin{aligned}
&P(c_{u,i} = c \mid z_{u,i} = k, t_{u,i} = t, c_{-u,i}, z_{-u,i}, t_{-u,i}, .) \\
&\quad \propto \frac{N_u^{(c)} + \gamma}{N_u^{(.)} + \mid C \mid \gamma} \cdot \frac{N_c^{(k)} + \alpha}{N_c^{(.)} + \mid K \mid \alpha} \cdot \\
&\qquad \frac{N_{c,k}^{(t)} + \epsilon}{N_{c,k}^{(.)} + \mid T \mid \epsilon}
\end{aligned} \tag{8}$$

$$
\begin{aligned}
&P(z_{u,i} = k \mid c_{u,i} = c, t_{u,i} = t, c_{-u,i}, z_{-u,i}, t_{-u,i}, w, .) \\
&\quad\propto \frac{N_c^{(k)} + \alpha}{N_c^{(.)} + \mid K \mid \alpha} \cdot \frac{N_{c,k}^{(t)} + \epsilon}{N_{c,k}^{(.)} + \mid T \mid \epsilon} \cdot \\
&\quad\frac{\prod_{v=1}^{M} \prod_{q=0}^{N_{u,i}^{(v)}-1} (N_k^{(v)} + q + \beta)}{\prod_{q=0}^{N_{u,i}^{(.)}-1} (N_k^{(.)} + q + \mid M \mid \beta)}
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&P(d_{m_{u,i},m_{v,j}} = 1, u, v \mid z_{u,i} = k, z_{v,j} = k, c_{u,i} = c, \\
&\quad c_{v,j} = c, s_u, v = s, c_{-u,i}, z_{-u,i}, c_{-v,j}, z_{-v,j}, .) \\
&\quad\propto \frac{N_c^{(k)} + \alpha}{N_c^{(.)} + \mid K \mid \alpha} \cdot \frac{N_u^{(c)} + \alpha}{N_u^{(.)} + \mid C \mid \gamma} \cdot \\
&\quad\frac{N_v^{(c)} + \alpha}{N_v^{(.)} + \mid C \mid \gamma} \cdot \frac{N_{u,v}^{(s)}}{N_{u,v}^{(.)} + \mid C \mid \gamma} \cdot
\end{aligned}
\tag{10}
$$

Furthermore, we can estimate the user–user topical influence strength. Suppose $\tau_{k,t}(u, v)$ represent the influence strength from user u to user v on topic k, which satisfy that $\tau_t(u, v) = \sum_{k=1}^{|K|} \tau_{k,t}(u, v)$. Thus, user–user topical influence strength can be estimated by

$$
\begin{aligned}
\tau_{k,t}(u, v) = \sum_{m_{u,i}=1}^{M_{u,t}} & P(d_{m_{u,i},m_{v,j}} = 1, u, v \mid z_{u,i} \\
&= k, z_{v,j} = k, c_{u,i} = c, c_{v,j} = c, \\
&\quad s_u, v = s, c_{-u,i}, z_{-u,i}, c_{-v,j}, z_{-v,j}, .)
\end{aligned}
\tag{11}
$$

where $N_u^{(c)}$ denotes the number of messages of user u assigned to community c; $N_c^{(k)}$ is the number of posts assigned to community c and generated by topic k; $N_{c,k}^{(t)}$ denotes the number of times that time stamp t is generated by community c and topic k; $N_{u,i}^{(v)}$ is the number of times word v occurs in the message $m_{u,i}$; $N_k^{(v)}$ denotes the number of times words v is assigned to topic k. Marginal counts are represented with dots, e.g., $N_{c,k}^{(.)}$ denotes the total number of time stamp generated by community c and topic k.

4.4 Time complexity

We now analyze the time complexity of the inference algorithm. In each iteration, the latent variables are sampled. Updates are performed by iterative sampling each latent variable from its conditional distribution, given the current values of the other variables. Since all the counters (e.g., $N_c^{(k)}$) involved in the previous equations can be caught and updated in constant time for each $c_{u,i}$ being sampled. Thus, sampling all c takes linear time w.r.t. the number of messages. Then, sampling all z is linear in the number of words. Overall, the inference algorithm takes linear time in the amount of data.

5 PRIN architecture

We propose a component-based architecture for the SIA performed. Figure 3 contains the diagram depicting the architecture of the system. We apply the discovered influence on user behavior prediction. The system consists of three main components:

- *Data collection* This module is in charge of compiling the user's data from a social networking service for getting the input to our system. The data depends on the social network such as structural information (friendship links in case of Facebook, following links in case of Twitter or Instagram, connections between users in case of LinkedIn), diffusion information (posts, tweets, messages). It continuously crawls a social network. Then, it stores the information gathered by the crawler in a Graph Database (Miller 2013).
- *Machine learning pipeline* This module generates a Bayesian network that describes the social influence process inside a social network. The stages are described below:

  1. *Data ingestion* The social network data is obtained and imported from the Graph Database, and this is called raw data.
  2. *Data preparation* A series of transformations in the raw data is performed to obtain it in the correct form. The transformations involve filling missing values or removing duplicate records or normalizing. This is where the feature extraction, construction, and selection takes place too to get the data model.
  3. *Code the model* The structure of the Bayesian network is built from a social network. It takes as input the number of users, number of topics, number of communities.
  4. *Train, evaluate, and tune the model* The data model is split into subsets of data to train the model, test it, and further validate how it performs against new data. The train set is used to calculate the parameters (probability distributions of the random variables). Then, the performance of the model is evaluated using the test and validation subsets of data to understand how accurate the prediction is. This is an iterative process until the model is calibrated.
  5. *Model deployment* Once the chosen model is produced, it is stored and embedded in decision-making frameworks as a part of an analytics solution.
  6. *Manage the model and version* The model is continuously monitored to observe how it behaved and calibrated accordingly to new data.

- *User interface* The user interface is composed of an inference algorithm that uses the model to answer queries and a visualization tool for the results.

6 Experimental evaluation

We performed extensive experiments with real-world data sets to quantitatively evaluate the performance of PRIN. The empirical evaluation is divided into multiple stages. First, we evaluated the PRIN's content component capability to extract topics. Moreover, we tested the PRIN's user component effectiveness for community detection. Then, we evaluated the PRIN's network component, where diffusion-links prediction performance was reported. At last, we performed a parameter analysis over the whole system for diffusion-links predicting.

6.1 Evaluation metrics

We used case studies to demonstrate further the effectiveness of our proposed model in real social networks using the following performance metrics.

- *Log-Likelihood* The log-likelihood is, as the term suggests, the natural logarithm of the likelihood. The likelihood measures the probability of the observed data, given the model, i.e., how well a model fits the observed data. The highest the likelihood, the better the model for the given data. For many applications, the log-likelihood is more convenient to work with (Koller et al. 2009). This is because we are generally interested in where the likelihood reaches its maximum value: the logarithm is a strictly increasing function, so the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques.
- *Perplexity* It is also a measure of model quality, and Natural Language Processing is often used as "perplexity per number of words." It describes how well a model predicts a sample, i.e., how much it is "perplexed" by a sample from the observed data. A lower perplexity score indicates better generalization performance. The perplexity of a discrete probability distribution (Brown et al. 1992) is calculated as follows:

$$2^{H(P)} = 2^{\sum_x P(x) \log_2 P(x)}, \tag{12}$$

 where $H(P)$ is the entropy of the distribution $P(x)$, and x is a random variable over all possible events.

- *AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve* AUC–ROC curve is a performance measurement for the classification problem at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. The curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The area under the curve is used to give a score to the model. If the area under the curve is 0.5, then the TPR is equal to the FPR, and the model is doing no better than random guessing.

Generally, a model with higher log-likelihood and lower perplexity is considered to be good. Then, a rough guide for evaluating a model with the AUC value is the traditional academic point system:

- .90–1 means the model is Excellent.
- .80–.90 means the model is Good.
- .70–.80 means the model is Fair.
- .60–.70 means the model is Poor.
- .50–.60 means the model Fails.

6.2 Baseline algorithms

We compared the results of our approach with previous work to evaluate PRIN. We choose baseline based on the following guidelines (1) They are state of the art to model heterogeneous networks; (2) They model diffusion prediction at the task level. Finally, we selected the baselines below and list our differences with them in Table 2.

- Poisson Mixed-Topic Link Model (PMTLM) proposed by Zhu et al. (2013) defines a generative process for both text and links between users.
- Followship-LDA (FLDA) proposed by Zhu et al. (2013) integrates both content topic discovery and social influence analysis in the same generative process. FLDA is a Bayesian generative model that extends Latent Dirichlet Allocation (LDA).
- COmmunity Level Diffusion Model (COLD) proposed by Hu et al. (2015) probabilistic generative model capturing influence between communities at different topics.
- Community Profiling and Detection (CPD) proposed by Cai et al. (2017) offers a generative model to formalize the concept of community profiling. They characterize a community in terms of its member users and both its internal content profile and external diffusion profile.
- Deepinf proposed by Qiu et al. (2018) deep neural network used for predicting social influence.

Table 2 Differences with baselines

Methods	Features			Diffusion factors			Task				
	Friends links	Diffusion links	Time	Ind.	Commu.	Topic	Topic extraction	Comm. Detec.	Diff. Pred	Ind. Profile	Comm. Profile
PMTLM (Zhu et al. 2013)		X				X	X		X		
FLDA (Bi et al. 2014)	X			X		X	X		X		
COLD (Hu et al. 2015)		X	X		X		X	X	X		
CPD (Cai et al. 2017)	X	X	X	X	X	X	X	X	X		X
Deepinf (Qiu et al. 2018)	X	X							X		
SVM (Fan et al. 2008)	X	X		X					X		
LR (Bishop 2006)	X	X		X					X		
PRIN	X	X	X	X	X	X	X	X	X	X	X

Fig. 4 Word cloud of the Amazon review corpus

- Support Vector Machine (SVM). We use the support vector machine with a linear kernel as a classifications model.
- Logistic Regression (LR). We also use the logistic regression (LR) to train a classification model.

6.3 Case study: Content component for topic modeling from Amazon reviews

The goal of this case study is to evaluate the performance of the PRIN content component for topic modeling with labeled data. We selected log-likelihood and perplexity as evaluation metrics.

6.3.1 Methodology

- *Data set* For evaluation of the PRIN's content component, we used Amazon products reviews data set. The data set is a list of 34,660 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database[1]. The data set includes basic product information, rating, review text, and more for each product. We decided to use this data set because it is labeled meaning by product; therefore, we can use this information for comparing our results. Besides, it is composed of short text like the ones that are shared in most social networks (Diao et al. 2012).
- *Data set preparation* For the aim of this study we extracted the following features: *categories* and *reviews. text*. Then, we only keep reviews with less than 200 words. The remained reviews followed these characteristics:

  - The average number of words in a review is 14.64.
  - The minimum number of words in a review is 5.
  - The maximum number of words in a review is 80.

[1] https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

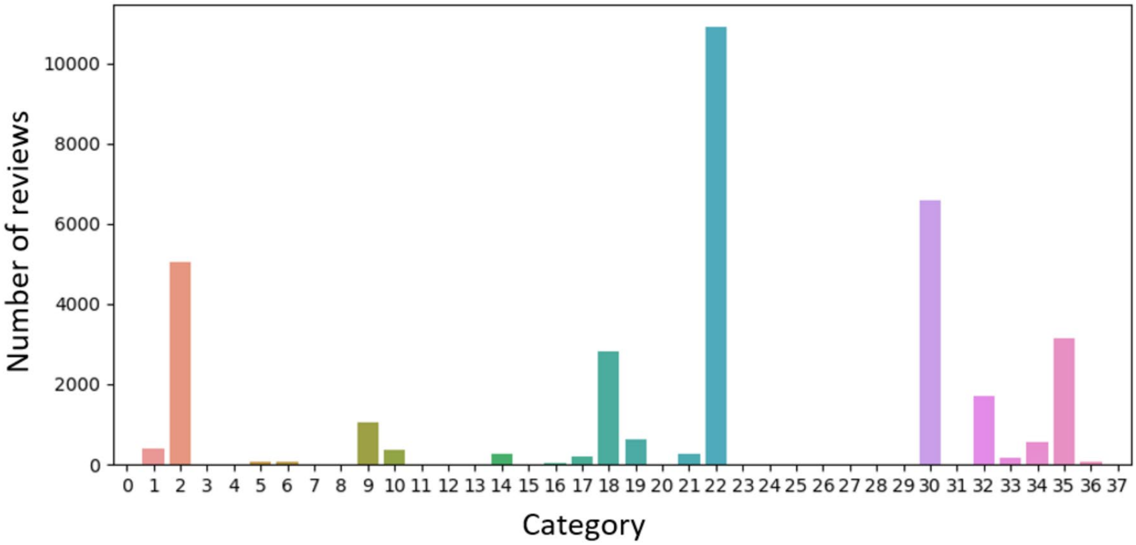


Fig. 5 Amazon reviews per categories

Figure 4 presents the word cloud of the corpus, and Fig. 5 shows the review distribution per category (38 total categories). It can be observed that there is an imbalanced problem in the data set; for solving this problem, we use the oversampling technique. Table 3 shows the name of the categories presented in Fig. 5. We processed the above data set following the practice in existing work for cleaning the data (Vega and Mendez-Vazquez 2016). However, we use a Bag of word (Goldberg 2017) approach for feature representation.

- *Experimental setup* Table 4 shows the values of the used parameters; batch size is the number of reviews to be used in each training chunk, alpha determines how often the model parameters should be updated, and epochs are the total number of training passes. After all, we were interested in determining what topic a given review is about, and we assigned this topic base on the highest percentage contribution of the topic in that review.

6.3.2 Results

Figure 6 shows the log-likelihood value under varying number of topics, and Fig. 7 shows the perplexity values under varying number of topics. Therefore, it can be observed that with $K = 15$, the model has a balance between perplexity and log-likelihood.

6.4 Case study: Content component for topic modeling from Twitter

The goal of this case study is to evaluate the PRIN content component performance for topic modeling in Tweets against other algorithms. However, in this case study, we do not have label data. Therefore, we evaluated the performance qualitatively following the practice of Chang et al. (2009) and quantitative using perplexity as a metric.

6.4.1 Methodology

- *Data set* The data set was crawled from Twitter. We started by choosing top popular users in specific topics such as Katty Perry, Donald Trump, Bill Gates, and Cristiano Ronaldo. Using these users as seed users, we crawled a network with about 3661 active users (Huberman et al. 2008). We extracted all the information posted by them from 12-14-2018 to 12-18-2018, which gave rise to 44209 following relationships, 56918 tweets, and 24649 retweets.
- *Data set preparation* For the aim of this study, we used only the users' tweets. Then, we processed the data set following the practice in existing work for cleaning the data (Vega and Mendez-Vazquez 2016). However, we use a Bag of words (Goldberg 2017) approach for feature representation.
- *Experimental setup* We used the same parameter setting as in Case study 6.3 presented in Table 4.

Table 3 Amazon reviews categories

Id	Categories
0	Electronics, iPad and Tablets, All Tablets, Fire Tablets, Tablets, Computers and Tablets
1	eBook Readers, Kindle E-readers, Computers and Tablets, E-Readers and Accessories, E-Readers
2	Electronics, eBook Readers and Accessories, Covers, Kindle Store, Amazon Device Accessories, Kindle E-Reader Accessories, Kindle (5th Generation) Accessories, Kindle (5th Generation) Covers
3	Tablets, Fire Tablets, Electronics, Computers, Computer Components, Hard Drives and Storage, Computers and Tablets, All Tablets
4	Tablets, Fire Tablets, Computers and Tablets, All Tablets
5	Amazon Devices and Accessories, Amazon Device Accessories, Power Adapters and Cables, Kindle Store, Kindle E-Reader Accessories, Kindle Paperwhite Accessories
6	Electronics, iPad and Tablets, All Tablets, Computers/Tablets and Networking, Tablets and eBook Readers, Computers and Tablets, E-Readers and Accessories, E-Readers, Used:Computers Accessories, Used:Tablets, Computers, iPads Tablets, Kindle E-readers, Electronics Features
7	Computers/Tablets and Networking, Tablets and eBook Readers, Electronics, eBook Readers and Accessories, eBook Readers
8	Fire Tablets, Tablets, Computers and Tablets, All Tablets, Electronics, Tech Toys, Movies, Music, Electronics, iPad and Tablets, Android Tablets, Frys
9	Kindle E-readers, Electronics Features, Computers and Tablets, E-Readers and Accessories, E-Readers, eBook Readers
10	Computers and Tablets, E-Readers and Accessories, eBook Readers, Kindle E-readers
11	Fire Tablets, Tablets, Computers and Tablets, All Tablets
12	Frys, Software and Books, eReaders and Accessories, Tablet Cases Covers, Tablet Accessories, Computer Accessories
13	Electronics, Categories, Streaming Media Players, Amazon Devices
14	Computers/Tablets and Networking, Tablets and eBook Readers, Computers and Tablets, Tablets, All Tablets
15	Amazon Device Accessories, Kindle Store, Kindle Touch (4th Generation) Accessories, Kindle E-Reader Accessories, Covers, Kindle Touch (4th Generation) Covers
16	Walmart for Business, Office Electronics, Tablets, Office, Electronics, iPad and Tablets, Windows Tablets, All Windows Tablets, Computers and Tablets, E-Readers and Accessories, E-Readers, eBook Readers, Kindle E-readers, Computers/Tablets and Networking, Tablets and eBook Readers, Electronics Features, Books and Magazines, Book Accessories, eReaders, TVs and Electronics, Computers and Laptops, Tablets and eReaders
17	Walmart for Business, Office Electronics, Tablets, Electronics, iPad and Tablets, All Tablets, Computers and Tablets, E-Readers and Accessories, Kindle E-readers, Electronics Features, eBook Readers, See more Amazon Kindle Voyage (Wi-Fi)
18	Electronics, Categories, Fire TV, Kindle Store
19	Electronics, Computers, Computer Accessories, Cases and Bags, Fire Tablets, Electronics Features, Tablets, Computers and Tablets, Kids' Tablets, Electronics, Tech Toys, Movies, Music, iPad and Tablets, Top Rated
20	Electronics, iPad and Tablets, All Tablets, Computers and Tablets, Tablets, eBook Readers
21	Kindle Store, Categories, eBook Readers and Accessories, Fire TV Accessories, Electronics, Power Adapters and Cables, Amazon Device Accessories, Power Adapters
22	Fire Tablets, Tablets, Computers and Tablets, All Tablets, Computers/Tablets and Networking, Tablets and eBook Readers
23	Categories, Streaming Media Players, Electronics
24	Computers and Tablets, Tablets, All Tablets, Computers/Tablets and Networking, Tablets and eBook Readers, Fire Tablets, Frys
25	Electronics Features, Fire Tablets, Computers and Tablets, Tablets, All Tablets, Computers/Tablets and Networking, Tablets and eBook Readers
26	Stereos, Remote Controls, Amazon Echo, Audio Docks and Mini Speakers, Amazon Echo Accessories, Kitchen and Dining Features, Speaker Systems, Electronics, TVs Entertainment, Clearance, Smart Hubs and Wireless Routers, Featured Brands, Wireless Speakers, Smart Home and Connected Living, Home Security, Kindle Store, Home Automation, Home, Garage and Office, Home, Voice-Enabled Smart Assistants, Virtual Assistant Speakers, Portable Audio and Headphones, Electronics Features, Amazon Device Accessories, iPod, Audio Player Accessories, Home and Furniture Clearance, Consumer Electronics, Smart Home, Surveillance, Home Improvement, Smart Home and Home Automation Devices, Smart Hubs, Home Safety and Security, Voice Assistants, Alarms and Sensors, Amazon Devices, Audio, Holiday Shop
27	Fire Tablets, Tablets, Computers and Tablets, All Tablets, Frys
28	TVs Entertainment, Wireless Speakers, Virtual Assistant Speakers, Featured Brands, Electronics, Amazon Devices, Home, Home Improvement, Home Safety and Security, Home Security, Alarms and Sensors, Smart Home and Home Automation Devices, Smart Hubs and Wireless Routers, Smart Hubs, Consumer Electronics, Voice-Enabled Smart Assistants, Smart Home and Connected Living, Home, Garage and Office, Smart Home, Voice Assistants, Surveillance, Home Automation, Speakers, Electronics Features, Holiday Shop, TV, Video and Home Audio, Internet and Media Streamers, Amazon Echo, Hubs and Controllers

Table 3 (continued)

Id	Categories
29	Chargers and Adapters, Computers and Accessories, Tablet and E-Reader Accessories, Amazon Devices and Accessories, Fire Tablet Accessories, Electronics, Power Adapters and Cables, Cell Phones, Amazon Device Accessories, Cell Phone Accessories, Cell Phone Batteries and Power, Tablet Accessories, Featured Brands, Kindle Fire (2nd Generation) Accessories, Kindle Store, Home Improvement, Fire (5th Generation) Accessories, Electrical, Amazon Devices, Home, Tablets and E-Readers, Cables and Chargers
30	Cases, Kindle Store, Amazon Device Accessories, Accessories, Tablet Accessories
31	Electronics, eBook Readers and Accessories, Power Adapters, Computers/Tablets and Networking, Tablet and eBook Reader Accs, Chargers and Sync Cables, Power Adapters and Cables, Kindle Store, Amazon Device Accessories, Kindle Fire (2nd Generation) Accessories, Fire Tablet Accessories
32	Electronics, Tablets and E-Readers, Tablets, Back To College, College Electronics, College Ipads and Tablets, Featured Brands, Amazon Devices, Electronics Deals, Computers and Tablets, All Tablets, Electronics Features, eBook Readers
33	Featured Brands, Electronics, Amazon Devices, Home, Home Improvement, Home Safety and Security, Home Security, Alarms and Sensors, Smart Home and Home Automation Devices, Mobile, Mobile Speakers, Mobile Bluetooth Speakers, Smart Hubs and Wireless Routers, Smart Hubs, Home, Garage and Office, Smart Home, Voice Assistants, Smart Home and Connected Living, Amazon Tap, Portable Audio, MP3 Accessories, Speakers, Amazon Echo, Electronics Features, TVs and Electronics, Portable Audio and Electronics, MP3 Player Accessories, Home Theater and Audio, Kindle Store, Frys, Electronic Components, Home Automation, Electronics, Tech Toys, Movies, Music, Audio, Bluetooth Speakers
34	Rice Dishes, Ready Meals, Beauty, Moisturizers, Lotions
35	Back To College, College Electronics, College Tvs and Home Theater, Electronics, Tvs and Home Theater, Streaming Devices, Featured Brands, Amazon Devices, Holiday Shop, Ways To Shop, TV and Home Theater, Streaming Media Players, All Streaming Media Players, TVs Entertainment, Video Games, Kindle Store, Electronics Features, Kids and Family, Fire TV
36	Electronics, Amazon Device Accessories, Kindle Store, Covers, Kindle E-Reader Accessories, Kindle DX (2nd Generation, Global Wireless) Accessories
37	Computers/Tablets and Networking, Tablet and eBook Reader Accs, Chargers and Sync Cables, Power Adapters and Cables, Kindle Store, Amazon Device Accessories, Fire Tablet Accessories, Kindle Fire (2nd Generation) Accessories

Table 4 Parameter setting for Case study 6.3

Parameter	Value
Batch size	128
Alpha	1
Epochs	500

6.4.2 Results

We reported the PRIN's content component capacity in extracting latent topics in the Twitter data set. For qualitative evaluation, we presented four example word clouds in Fig. 8.

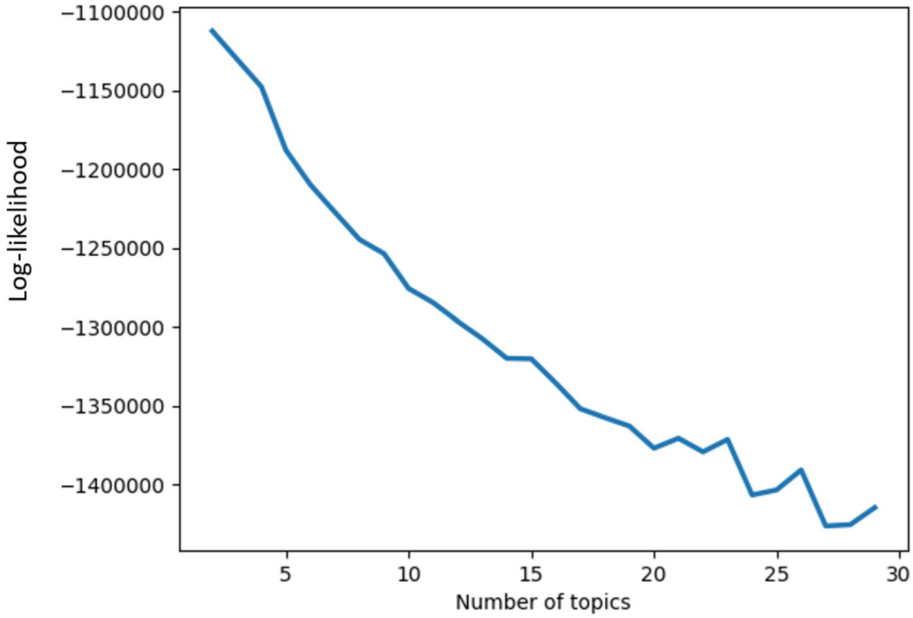


Fig. 6 Log-likelihood scores of PRIN's content component (the high, the better) from Case study 6.3

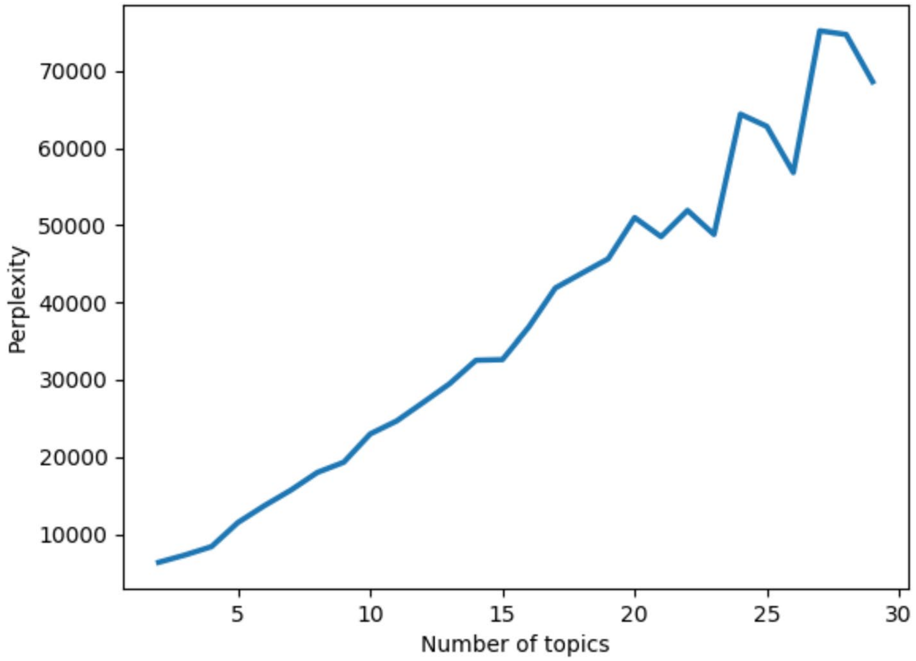
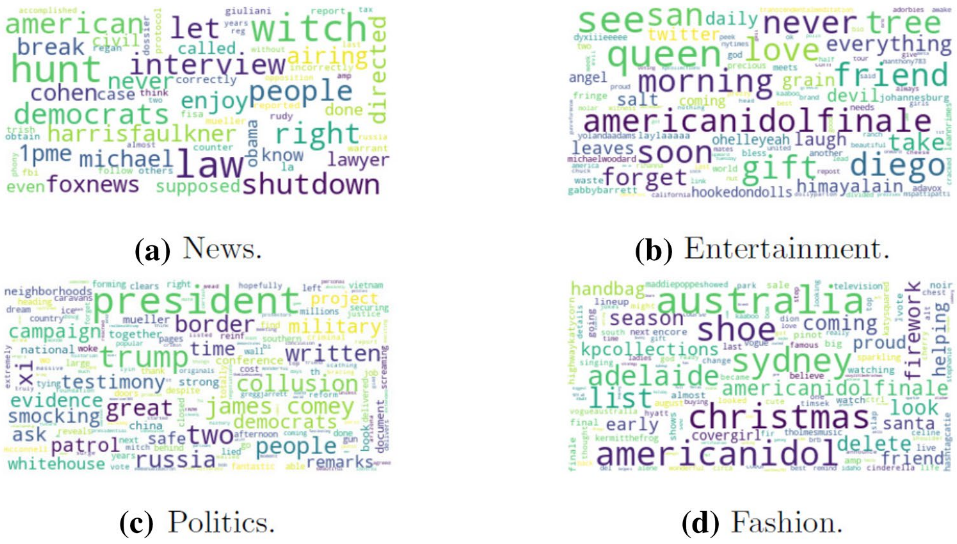


Fig. 7 Perplexity scores of PRIN's content component (the less, the better) from Case study 6.3



(a) News.

(b) Entertainment.

(c) Politics.

(d) Fashion.

Fig. 8 Word clouds of extracted topics

For quantitative evaluation, Fig. 9 shows the perplexity values under varying number of topics. It revealed that PRIN ($K = 100$) has the lowest perplexity, indicating the best topic discovery performance among all the competitors. Perplexity scores for COLD and PRIN were close, and both significantly outperform PMTLM.

6.5 Case of study: User component for community detection in Twitter

The goal of this case study is to evaluate the performance of the PRIN user component for community detection. For quantitative evaluation, due to the lack of ground truth for communities inside the network, we use link prediction, a widely used quantitative measurement in the mixed-membership community setting without community labels (Biswas and Biswas 2017). Link prediction is defined to estimate the probability of a link between two users. Moreover, as there is not a predefined threshold for link prediction, we used AUC as the evaluation metric.

6.5.1 Methodology

- *Data set* The Higgs Boson Twitter data set (De Domenico et al. 2013) was built after monitoring the spreading pro-

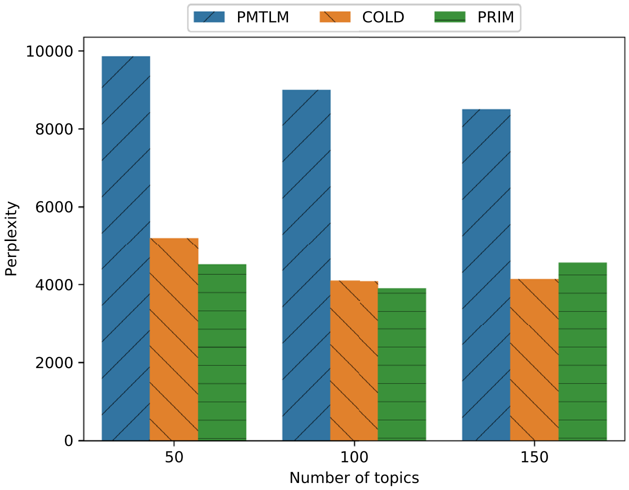


Fig. 9 Perplexity scores from Case study 6.4

cesses on Twitter before, during, and after the announcement of the discovery of a new particle with the features of the elusive Higgs Boson on July 4th, 2012. The data set contains three types of social interaction mentioning, retweeting, replying to existing "Higgs" related tweets and friends/followers social relationships among users involved in the above activities. The data set statistics are presented in Table 5. It is worth remarking that the user IDs have been anonymized, and the same user ID is used for all networks.

- *Data set preparation* For the aim of this study, we used only the followers' social relationships among users. We processed the data set as follows: for every pair of users, we generated a positive instance if there is a follower relationship; otherwise, we created a negative instance. Our target is to distinguish positive instances from negative ones.
- *Experimental setup* Table 6 shows the values of the used parameters.

Table 5 Data set statistics of Higgs Twitter Data set

Social network statistics	
Nodes	456,626
Edges	14,855,842
Nodes in largest WCC	456,290 (0.999)
Edges in largest WCC	14,855,466 (1.000)
Nodes in largest SCC	360,210 (0.789)
Edges in largest SCC	14,102,605 (0.949)
Average clustering coefficient	0.1887
Number of triangles	83,023,401
Fraction of closed triangles	0.002901
Diameter (longest shortest path)	9
90-percentile effective diameter	3.7
Retweet network statistics	
Nodes	256,491
Edges	328,132
Nodes in largest WCC	223,833 (0.873)
Edges in largest WCC	308,596 (0.940)
Nodes in largest SCC	984 (0.004)
Edges in largest SCC	3850 (0.012)
Average clustering coefficient	0.0156
Number of triangles	21,172
Fraction of closed triangles	0.0001085
Diameter (longest shortest path)	19
90-percentile effective diameter	6.8
Reply network statistics	
Nodes	38,918
Edges	32,523
Nodes in largest WCC	12,839 (0.330)
Edges in largest WCC	14,944 (0.459)
Nodes in largest SCC	322 (0.008)
Edges in largest SCC	708 (0.022)
Average clustering coefficient	0.0058
Number of triangles	244
Fraction of closed triangles	0.0001561
Diameter (longest shortest path)	29
90-percentile effective diameter	10

6.5.2 Results

Figure 10 shows the AUC values for the models. PRIN $(C = 100,\ K = 100)$ outperformed all other methods. Moreover, COLD and PRIN were significantly better than PMTLM.

6.6 Case study: Network component for influence prediction in Twitter

The goal of this case study is to evaluate the PRIN network performance for diffusion link prediction (retweets) on Twitter. For quantitative evaluation, we selected AUC as an evaluation metric, since there is not a predefined threshold for link prediction.

Table 6 Parameter setting for Case study 6.5

Parameter	Value
Batch size	128
Alpha	1
Epochs	350

6.6.1 Methodology

- *Data set* We used the Hibbs Boson Twitter network detailed in Case study 6.5.
- *Data set preparation* We processed the data set following the practice in existing work (Zhang et al. 2013, 2015).

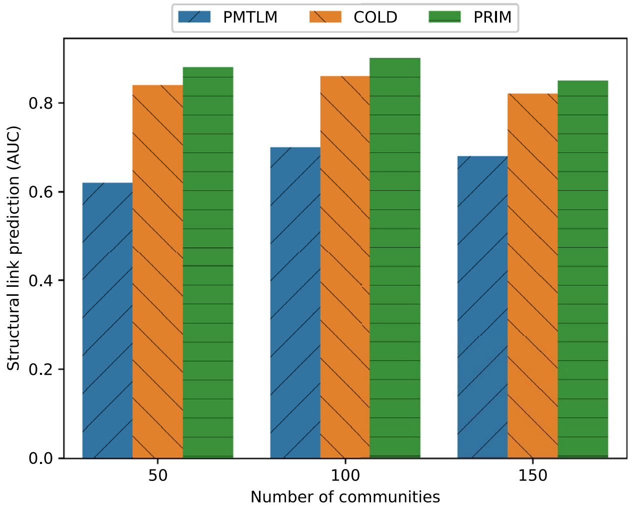


Fig. 10 Structural link prediction performance

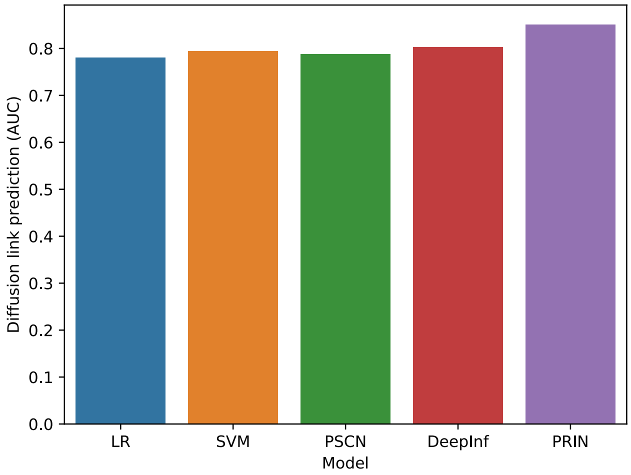


Fig. 11 Diffusion link prediction performance from Case study 6.6

Furthermore, for a user v who was influenced to perform a social action a at some timestamp t, we generate a positive instance. Moreover, if a user v was never observed to be active in our observation window, we create a negative instance. For instance, analyzing observation windows of 5 consecutive days, for a tweet a generated at timestamp t by a user v, we analyze if it is a retweet from another tweet. In this case, we generate a positive instance; otherwise, it will be considered as a negative instance in that observation window. Our target is to distinguish positive instances from negative ones in the observation window.

- *Experimental setup* We performed nested cross-validation over a windows time per user in the sampled social network of 100 users. The window size used for the experiments was of 5 days. The historical validation consists of the following:

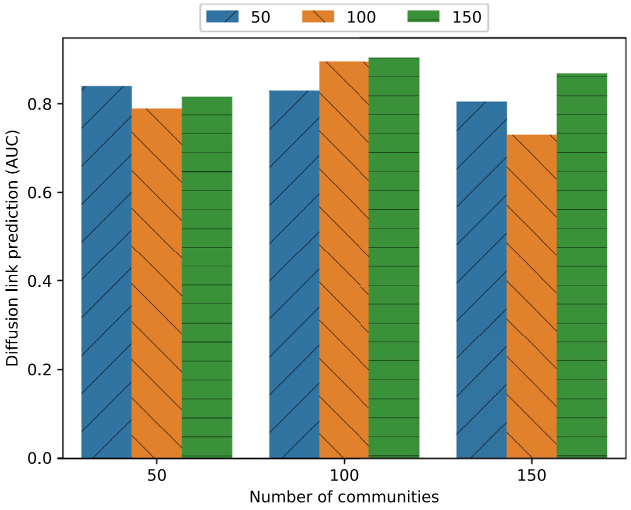


Fig. 12 The impact of model parameters C and K in the task of diffusion link prediction

1. Extract all the timeline information (observed variables) from the set of users: message, structural links, and diffusion links.
2. Split the timeline into K "folds" (sets) of the size according to the window size: the most recent information will be used as a test set, and the oldest will be used as a train set.
3. Calculate the CPDs using the train set for the K-fold.
4. Calculate the predictions from the test set, removing the feature of diffusion and set the rest of the variables as evidence.
5. Compare the prediction with the real labels calculating the AUC.
6. Repeat Step 3 to 5 for every fold, merging the previous train set K with the new one $K + 1$.
7. Get the average AUC for all the folds.

6.6.2 Results

We compared the prediction performance of all methods across the Higgs Boson data set; Fig. 11 presents the results. PRIN achieved significantly better performance over baselines in terms of AUC, demonstrating the effectiveness of our model.

6.7 Case of study: Parameter analysis

PRIN model is mostly affected by two parameters, i.e., a number of communities C and a number of topics K. We investigated how the prediction performance varies with these parameters.

6.7.1 Methodology

- *Data set* We used same Twitter data set as in Case study 6.4.
- *Data set preparation* We applied the same data preparation as in Case study 6.6.
- *Experimental setup* We used the same parameter values as in Case study 6.4 presented in Table 4.

6.7.2 Results

Figure 12 shows the AUC values of diffusion link prediction under different settings, such as a number of topics and a number of communities.

7 Discussion

This research was focused on the social influence's characterization problem. We proposed a probabilistic model called PRIN. PRIN was designed to reveal the rich spectrum of the online social influence process. Then, we validated each component with case studies. We discussed their main findings in this section.

In Case study 6.3, we evaluated content component with a labeled data set. We were expected the best performed with 38 numbers of topics, due to this value was the real number of categories labeled by Amazon. However, we discovered that the model performed well with 15 topics, with a low perplexity and high log-likelihood. This is due to the fact that some of the categories overlap each other, as follows:

- eBook Readers, Kindle E-readers, Computers and Tablets, E-Readers and Accessories, E-Readers.
- Electronics, eBook Readers and Accessories, Covers, Kindle Store, Amazon Device Accessories, Kindle E-Reader Accessories, Kindle (5th Generation) Accessories, Kindle (5th Generation) Covers.
- Electronics, iPad and Tablets, All Tablets, Computers/Tablets, and Networking, Tablets and eBook Readers, Computers and Tablets, E-Readers and Accessories, E-Readers, Used:Computers Accessories, Used:Tablets, Computers, iPads Tablets, Kindle E-readers, Electronics Features.
- Electronics, iPad and Tablets, All Tablets, Computers and Tablets, Tablets, eBook Readers.

Moreover, the results of the Content component's performance in Case study 6.4 were the lowest perplexity, against the other algorithms for topic modeling in Tweets. This means that the component is doing well compared with the state-of-the-art algorithms. Furthermore, Case study 6.5 presented the AUC values for structural link prediction. PRIN ($C = 100$, $K = 100$) outperformed the others algorithms with a sample of 100 users. This can be attributed to their lack of representation of the user's profile. Besides, as shown in Fig. 11 from Case study 6.6, our model consistently outperforms all the baselines, thanks to our modeling various diffusion factors and heterogeneous user links, in contrast with the baselines in Table 2. Then, in Case study 6.7 is observed that the model is highly dependent on the values of the K and C parameters. They together exert influence on diffusion prediction accuracy. The performance is stable under a broad range of parameter settings, indicating little tuning is required in actual deployment.

Generally, the quality of the predictions depends on two things: the degree to which the original model accurately reflects real-world situations, and the amount of data you provide. One limitation of our research was the sample size, which was too small for large-scale social networks. However, from the results of those limited number of users, it seems that the model can reflect the social influence process with high values of AUC.

Moreover, our research only focuses on one type of social networks such as Twitter[2], the results can be generalized to any other social networks such as Diggs[3], Reddit[4]. On Twitter, the user action is defined as whether a user posts/re-tweets a blog on a specific topic. On Digg and Reddit, the action is defined as whether a user submits/votes a story on a topic.

8 Conclusion and future work

In this paper, we study the social influence modeling problem. We introduce a Probabilistic Reasoning System for Social Influence Analysis (PRIN) to describe the problem using a graphical probabilistic model. The model leverages both heterogeneous link information, time, and content associated with each user in the network to mine topic-level influence strength in online social networks.

The extensive experimental results demonstrate the effectiveness of our model. PRIN achieved the best performance in the tasks of link prediction and text perplexity among several competitors.

The existing models fail to give a complete view of social influence interaction dynamics. The proposed model provides a visual and more in-depth understanding of user interactions, which can intuitively capture the process that drives social influence.

[2] http://www.twitter.com, a microblogging system.

[3] http://www.digg.com, a social news sharing and voting website.

[4] https://www.reddit.com/r/socialmedia/, a social sharing website.

For future work, it is known that users' behaviors are distributed in different networks, we are interested in merging the information from various social networks and leverage the correlation between them to better performance of the influence learning. Then, since the proposed model is generative, this means that arbitrary queries can be answered, not only about the probability of a diffusion. It will be interesting to explore other questions such as community memberships and interest detection in a topic. Another interesting issue is to employ semisupervised learning to incorporate user feedback into our approach. Then, the content component could be replaced for a supervised algorithm which could improve the accuracy of the predictions. Besides, since social networks are growing in size, it will be interesting to explore a distributed implementation of the proposed architecture to handle large-scale social networks.

Acknowledgements The authors would like to thank CONACYT for the financial support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

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

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Biswas A, Biswas B (2017) Community-based link prediction. Multimedia Tools Appl 76(18):18619–18639

Brown JJ, Reingen PH (1987) Social ties and word-of-mouth referral behavior. J Consum Res 14(3):350–362

Brown PF, Pietra VJD, Mercer RL, Pietra SAD, Lai JC (1992) An estimate of an upper bound for the entropy of English. Comput Linguist 18(1):31–40. http://dl.acm.org/citation.cfm?id=146680.146685

Cai H, Zheng VW, Zhu F, Chang KCC, Huang Z (2017) From community detection to community profiling. Proc VLDB Endow 10(7):817–828

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems, pp 288–296

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, ACM, pp 160–168

De Domenico M, Lima A, Mougel P, Musolesi M (2013) The anatomy of a scientific rumor. Sci Rep 3:2980

Diao Q, Jiang J, Zhu F, Lim EP (2012) Finding bursty topics from microblogs. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers-volume 1, Association for Computational Linguistics, pp 536–544

Doucet A, De Freitas N, Murphy K, Russell S (2000) Rao-Blackwellised particle filtering for dynamic Bayesian networks. In: Proceedings of the sixteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp 176–183

Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. In: ACM SIGCOMM computer communication review, ACM, vol 29, pp 251–262

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9(Aug):1871–1874

Fei H, Jiang R, Yang Y, Luo B, Huan J (2011) Content based social behavior prediction: a multi-task learning approach. In: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, pp 995–1000

Franks H, Griffiths N, Anand SS (2013) Learning influence in complex social networks. In: Proceedings of the 2013 international conference on autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 447–454

Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Hum Lang Technol 10(1):1–309

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

Granovetter MS (1977) The strength of weak ties. In: Social networks, Elsevier, pp 347–367

Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: a survey. ACM Sigmod Record 42(2):17–28

Han M, Li Y (2018) Influence analysis: a survey of the state-of-the-art. Math Found Comput 1(3):201–253

Hu Z, Wang C, Yao J, Xing E, Yin H, Cui B (2013) Community specific temporal topic discovery from social media. ArXiv preprint arXiv:13120860

Hu Z, Yao J, Cui B, Xing E (2015) Community level diffusion extraction. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, ACM, pp 1555–1569

Huberman BA, Romero DM, Wu F (2008) Social networks that matter: Twitter under the microscope. ArXiv preprint arXiv:08121045

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, ACM, pp 137–146

Koller D, Friedman N, Bach F (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge

Li M, Wang X, Gao K, Zhang S (2017) A survey on information diffusion in online social networks: models and methods. Information 8(4):118

Liu L, Tang J, Han J, Yang S (2012) Learning influence from heterogeneous social networks. Data Min Knowl Disc 25(3):511–544

March JG (1955) An introduction to the theory and measurement of influence. Am Polit Sci Rev 49(2):431–451

Miller JJ (2013) Graph database applications and concepts with Neo4j. In: Proceedings of the southern association for information systems conference, Atlanta, GA, USA, vol 2324

Nallapati RM, Ahmed A, Xing EP, Cohen WW (2008) Joint latent topic models for text and citations. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 542–550

Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Pálovics R, Benczúr AA, Kocsis L, Kiss T, Frigó E (2014) Exploiting temporal influence in online recommendation. In: Proceedings of the 8th ACM conference on recommender systems, ACM, pp 273–280

Pezzoni F, An J, Passarella A, Crowcroft J, Conti M (2013) Why do i retweet it? An information propagation model for microblogs. In: International conference on social informatics, Springer, pp 360–369

Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J (2018) DeepInf: social influence prediction with deep learning. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 2110–2119

Rashotte L (2007) Social influence. The Blackwell Encyclopedia of Sociology, pp 4426–4428

Saito K, Nakano R, Kimura M (2008) Prediction of information diffusion probabilities for independent cascade model. In: International conference on knowledge-based and intelligent information and engineering systems, Springer, pp 67–75

Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268

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, ACM, pp 807–816

Travers J, Milgram S (1967) The small world problem. Phychol Today 1(1):61–67

Vega L, Mendez-Vazquez A (2016) Dynamic neural networks for text classification. In: 2016 international conference on computational intelligence and applications (ICCIA), IEEE, pp 6–11

Wen Z, Lin CY (2010) On the quality of inferring interests from social neighbors. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 373–382

Weng J, Lim EP, 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, ACM, pp 261–270

Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Proceedings of the 19th international conference on World wide web, ACM, pp 981–990

Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput Surv (CSUR) 45(4):43

Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: 2010 IEEE international conference on data mining, IEEE, pp 599–608

Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z (2010) Understanding retweeting behaviors in social networks. In: Proceedings of the 19th ACM international conference on information and knowledge management, ACM, pp 1633–1636

Zhang J, Liu B, Tang J, Chen T, Li J (2013) Social influence locality for modeling retweeting behaviors. In: Twenty-third international joint conference on artificial intelligence

Zhang J, Tang J, Li J, Liu Y, Xing C (2015) Who influenced you? Predicting retweet via social influence locality. ACM Trans Knowl Discov Data (TKDD) 9(3):25

Zhu Y, Yan X, Getoor L, Moore C (2013) Scalable text and link analysis with mixed-topic link models. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 473–481