Knowledge-Based Systems 000 (2017) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Predicting information diffusion probabilities in social networks: A Bayesian networks based approach

Devesh Varshney[a,b], Sandeep Kumar[a,*], Vineet Gupta[b]

[a] Department of Computer Science and Engineering, IIT Roorkee, India
[b] Adobe Research Labs, Bangalore, India

ARTICLE INFO

Article history:
Received 28 April 2017
Revised 28 June 2017
Accepted 1 July 2017
Available online xxx

Keywords:
Social network analysis
Information diffusion
Diffusion network
Bayesian network modeling
Diffusion probability

ABSTRACT

In past few years, social networking has significantly contributed to online presence of users. These social networks are hosts to a number of viral phenomena. This has fetched a lot of attention from various researchers and marketers all over the world. Major portion of the studies done in the field of information diffusion through social networks has focused on the problem of influence maximization. These methods demand the diffusion probabilities associated with the links in the social networks to be provided as inputs. However, the problem of computing these diffusion probabilities has not been as widely explored as the problem of influence maximization. In this paper, we tackle the problem of predicting the probabilities of diffusion of a message through the links of a social network. This paper presents a Bayesian network based approach for solving the aforesaid problem. In addition to the features related to the social network, this machine learning based Bayesian framework utilizes user interests and content similarity modeled using the latent topic information. We evaluate the proposed method using the data obtained from the well-known social network platform - Twitter.

1. Introduction

In present scenario, social networking and micro-blogging sites have become dynamic and widely used media for communication. Using these sites, people share information on various topics through which they express their likes and interests. As a result, social networking sites like Facebook[1] and Twitter[2] have shown a tremendous potential to make content viral, instantly, for example, Twitter during the United States presidential election in 2008 [1] and Facebook during the 2010 Arab spring [2]. Today, social networks are hosts to a number of viral phenomena: breaking news propagation, information dissemination during emergency, marketing campaigns, etc. [3,4]. This potential has caught the eyes of researchers as well as marketers, prompting them to focus on the word-of-mouth marketing strategy using these social platforms.

Studying and modeling the information propagation through social networks is important in making effective use of social platforms. It is not only helpful in understanding how information is diffused in the online social networks, but it can also be leveraged for solving a number of problems like influence maximization, personalized recommendation systems, trending topics detection, trust propagation, feed ranking in social networking sites, ad delivery, computing diffusion centrality measures in social networks [5–11].

Several researchers have studied the information diffusion through online social networks in the past. Most of the works in this area are based on the probabilistic models of information diffusion through networks, namely independent cascade (IC) and linear threshold (LT) [12]. These works assume the information diffusion probability (IDP) for each link in the network to be given as input. Some research works [7,13,14] have addressed the problem of predicting information diffusion, i.e., whether a user will retweet or not. On the other hand, the problem of computing IDP values has not been widely explored. Recently, some works [8,15–17] have reported the study of influence computation in social networks. Some of these works [8,13,16] use only network dynamics for creating solution models. However, the use of network dynamics alone cannot accurately capture the user interests and other relevant features [18]. Studies show that information dissemination processes are homophily-driven [19] and message propagation occurs more frequently between users having common interests [13]. Recently, Romero et al. [20] have proved that the textual

* Corresponding author at: Department of Computer Science and Engineering, IIT Roorkee, India.
E-mail addresses: devu.var@gmail.com (D. Varshney), sgargfec@iitr.ac.in, sandeepkumargarg@gmail.com (S. Kumar), vineetgupta10@gmail.com (V. Gupta).

[1] www.facebook.com
[2] www.twitter.com

http://dx.doi.org/10.1016/j.knosys.2017.07.003
0950-7051/

content plays a significant role in information propagation through social networks.

Taking these findings into consideration, we present a novel approach to predict the information diffusion probabilities of a message through various links in a social network. The approach uses textual content of messages, diffusion history of the network, and the network and user characteristics to build a Bayesian network model. The method exploits latent information extracted from the textual content to compute various features like user similarity, content similarity and user interests as described in Section 3.2. Given a social network and its diffusion history, the proposed approach finds the probabilities associated with the links of the social network for diffusion of a given message. Following are the contributions of this work:

- We analyze different factors which affect the IDPs and present a generic approach to compute these as features using the data available from online social networks.
- This work investigates the dependencies that exist among these features and how can we leverage this information to build a reliable model.
- This paper presents a Bayesian network based approach to compute the IDP values for different contents in a social network.
- We compare the previous studies in this field and demonstrate how latent topic information obtained from diffused message contents can be utilized to improve the state-of-the-art methods.
- The experimental evaluation of the proposed approach on real-world data demonstrates the effectiveness and efficiency of the method.

In this paper, we study the literature on information diffusion process from sociology and computer-science background and then perform experimental study to get better insights in order to answer the following research questions:

- **RQ1**: *What factors related to network settings, user characteristics, information source and message content affect the IDP values? How one can extract information out of the data available from online social networks to quantify these factors as features?*
- **RQ2**: *Whether these features are independent or there exist some dependencies among these and how one can formulate these dependencies?*
- **RQ3**: *How to design a model which can leverage these features and dependencies to reliably predict the IDP values in a social network for a given message?*

The rest of this paper is organized as follows. Section 2 provides background and discusses relevant works studying information diffusion. In Section 3, we present an approach for the computation of IDPs. Section 4 presents the experiments performed to evaluate the proposed method and the comparative analysis of the proposed approach with existing approaches. Section 5 presents the discussion of the experimental study. Finally, Section 6 concludes the paper with discussion of the proposed approach and scope for future works.

2. Related work

A number of researchers in the past have studied the process of information diffusion through social networks. These studies can be largely categorized into two parts. First, a significant amount of these works focus on addressing the problem of *Influence Maximization*. The goal of *Influence Maximization* is to find a seed set of users who can trigger cascades for maximizing the spread of an idea or opinion in the social network. For the first time, Domingos and Richardson [5,21] addressed this problem. They proposed a probabilistic model using Markov random fields. Later on, Kempe et al. [12] modeled the same as discrete optimization problem and proved its NP-hardness for two basic diffusion models namely–independent cascade and linear threshold. They also proposed a greedy approximation algorithm for solving the same. Most of the works studying the problem of *Influence Maximization* assume the IDPs of the links are given as inputs. In this work, instead we address the problem of predicting these probabilities. Second, some of these works aim to study the prediction of information diffusion, which includes determining whether a link will be active or not and what IDPs to assign to such links in social networks. Fei et al. [13] used a multi-task learning approach to predict a user's response (like or comment) to a post of her friend. They used features representing content similarity and user interests in their approach. Lin et. al [22]. proposed a probabilistic model TIDE (text-based information diffusion and evolution), to track the evolution of a topic with time in social communities and its diffusion paths. Their model extract features from text of posts and captures implicit features using Gaussian random field. But, this model ignores the features related to social connections.

Zhu et al. [15] studied the retweeting behavior of users and presented a logistic regression model to predict the retweeting probability of the incoming tweet by the target user. They used features related to the network such as the numbers of friends, followers, mutual friends, mutual followers, mutual mentions, mutual retweets, and the status count of the tweet author to capture the relationship among users. The model also takes into account the timing of tweets. They modeled content influence using URLs, mentions, and hash-tags in the tweet. The work captured topic similarity between the incoming tweet and the tweets of the user as the cosine similarity of their term frequency vectors. However, term frequency vectors are very sparse due to diverse use of vocabulary of interacting users.

Kuo et al. [14,23] have addressed the diffusion prediction on novel topic problem to predict both cross-topic-observed and unobserved diffusions. They used the latent information present in the posts to model a signature for each of the topics defined in the topic set and to model the user preferences towards these topics. These methods consider the feature related to the topic of the message but not the features related to the content of that message. Thus, the diffusion is specific to a topic irrespective of what is exact content of the message. However, we argue that the content of a message is also significant in the process of diffusion, thus we consider the message content for feature modeling. Varshney et al. [24] also addressed the problem of detecting the links which are active in the propagation of a given message through a social network. The works discussed above do not provide the probability of diffusion of the message through the link as the models used are classification based and not the probabilistic ones.

Similar to our work, the works in [8,16] also explored the problem of finding information diffusion probabilities. Saito et. al [16]. presented a method for predicting information diffusion probabilities for independent cascade (IC) model. This work defined the likelihood of multiple diffusion episodes and applied the Expectation Maximization (EM) algorithm to solve it. But this method is not scalable to huge datasets because EM algorithm has to update the diffusion probability of each link in each iteration. Goyal et al. [8] proposed models for learning the probabilities of influence between users with the focus on linear threshold (LT) model. They proposed various models like static model of probabilities, continuous time model and discrete time model. They also presented a technique for predicting the time by which a user is expected to perform the action. Due to the assumptions taken in their method such as a user performs an action at most once and the influence graph is DAG, their method is not suitable for tweet-retweet networks. Both of these works [8,16] ignore the content of posts to

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

Table 1
Summary of information diffusion studies w.r.t the information used for feature modeling and output task type.

Works								
Features	Fei et al. [13]	Zhu et al. [15]	Kuo et al. [14,23]	Saito et al. [16]	Goyal et al. [8]	Zhang et al. [7]	Jiang et al. [17]	Proposed
Diffusion history	✓	–	✓	✓	✓	✓	✓	✓
Network connection	–	✓	✓	–	✓	✓	–	✓
User activities	–	–	✓	–	–	–	✓	✓
Time delay	–	✓	–	–	✓	–	–	–
Hashtags, URLs, mentions	–	✓	–	–	–	–	–	–
Topic information	–	–	✓	–	–	–	–	✓
Content based similarities	✓	✓	–	–	–	✓	✓	✓
Output type (P/C) (**P**robabilistic/**C**lassification)	C	C	C	P	P	C	P	P

formulate the solution to the problem of finding probabilities of information diffusion while we consider the message content plays significant role in information diffusion.

Recently, Zhang et al. [7] have modeled the retweet prediction task using probabilistic matrix factorization technique. They modeled social contextual relationships and message semantics in latent feature space. Also, Jiang et al. [17] have analyzed the fundamental factors that affect retweetability of a tweet and then used one-class collaborative filtering to predict user's retweeting behavior. They also quantified the personal preferences of a user and the amount of social influence between users. Finally, they define the retweetability score by using a linear ensemble from user's preferences and social influence. These methods use only message semantic information and user preferences modeled in latent feature space and ignore the information related to network dynamics.

In this paper, the proposed method predicts the information diffusion probabilities by considering the content of posts, user characteristics, and the diffusion history of the network. It helps in modeling the information diffusion probabilities for the messages related to different topics differently. To the best of our knowledge, this is a novel work as no other work has leveraged all the information related to network dynamics, diffusion history, message semantics and user preferences in a unified model to solve the problem of computing diffusion probabilities.

Table 1 summarizes the surveyed methods according to following criteria that allows a quick comparison: *(i)* what information is used for feature modeling, and *(ii)* whether the method performs classification task to predict if the user/ link will be active or it performs a probabilistic modeling to predict the IPDs. It should be noted that the table is not presented with the intention of expressing any preference, rather to present an overall comparison.

3. Bayesian network based approach to predict information diffusion probabilities

In this section, we present an approach for the prediction of information diffusion probability through various links of a social network. Block diagram shown in Fig. 1 represents various steps involved in the proposed solution framework. Initially, we discuss the data collection process (Section 3.1) followed by the details of feature extraction process (Section 3.2) to compute the features related to content of message, user characteristics and network topology. Next, we discuss how to learn Bayesian network and related parameters (Section 3.3). Finally, Section 3.3.3 presents the method to test the proposed model.

3.1. Data collection

As there is no publicly available data corpus comprising of tweet contents, retweet information, and user network information for studying information diffusion in the network, we created our own. We collected over one million publicly available tweets and retweets from Twitter. We used the REST APIs[3] provided by Twitter for this purpose. Since the API provides only a small fraction of tweets randomly, we collected the retweets of all the tweets separately. The collected tweets belong to topics of shopping, politics, social media, and festive season. The topics and the corresponding keywords used for searching the tweets are shown in Table 2. We also collected the user profile information including name, followers count, friends count, status count and location.

Table 2
Keywords related to topics used for data collection.

Topic category	Keywords
Shopping	shopping, sales, deals, buy, gosf[a]
Politics	politics, democracy, elections, delhipoll, aap[b], bjp, congress
Social media	socialmedia, facebook, twitter, whatsapp, smm, linkedin
Festive season	holidays, chrismas, xmas, newyear, gifts, presents

[a] gosf is an abbreviation for great online shopping festival.
[b] aap, bjp and congress are three popular political parties in India.

Table 3
Number of tweets and retweets in the collected dataset.

Topic category	No. of Tweets	No. of Retweets
Shopping	227,203	39,087
Politics	240,710	59,612
Social media	219,099	40,328
Festive season	181,404	35,649

The resulting dataset contains over 70 thousand users (nodes) and over 175 thousand retweets (links). Table 3 lists the topic-wise distribution of tweets. Here, the number of retweets indicate the total number of tweets collected which are retweets of some tweet. Thus, if a tweet gets retweeted n times, it is counted n times.

For our analysis, we use the retweet graph created using collected nodes, tweets and retweets. In order to construct the training corpus, for each topic, we consider the above mentioned retweet links as positive instances. We also sample negative instances equal to the number of positive instances for each topic. The negative instances are sampled by randomly selecting absence of retweet links of that topic between the nodes in the graph. This process is similar to the one followed by Kuo et al. [14].

3.2. Feature extraction

In this section, we describe our approach of selecting features for designing an efficient Bayesian learning model to compute diffusion probabilities. An instance of the corpus can be considered as a tuple (s, d, m, t), where s is the source of diffusion, i.e., user who posted the message, d is the destination of diffusion, i.e., user

[3] https://dev.twitter.com/docs/api/1.1 .

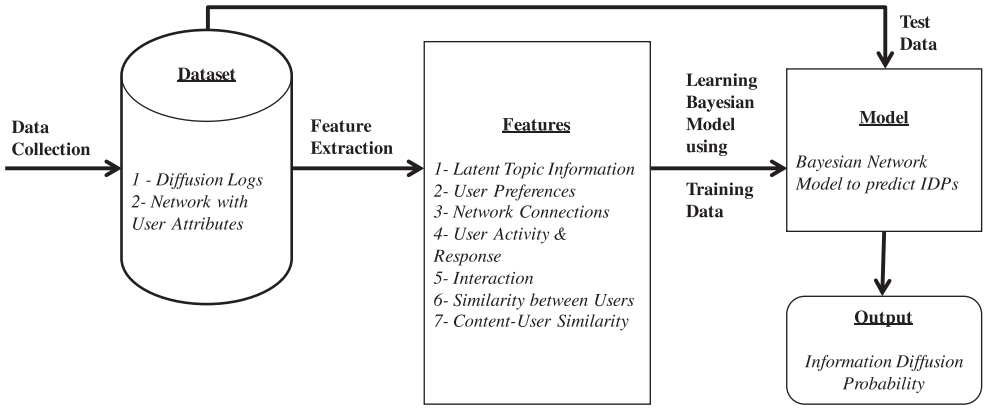

Fig. 1. Bayesian network based solution framework to predict information diffusion probabilities.

who reacted/ responded to the message, m is the message content of the diffusion and t is the topic category of the message m, like the ones defined in Table 2. We consider the diffusion edges to be directed. Following is an example of diffusion instance:

(anon1, anon2, "Important News: EC has said that voting will end only when the last voter in the queues have voted. No timeline fixed for poll end in Delhi", Politics).

In order to build the model, we compute various features related to the elements involved in the diffusion instance (s, d, m, t) as described in following sub-sections.

3.2.1. Latent Topic Information (LTI)

Topic modeling identifies the distribution of latent topics in the text, which is useful in modeling the interest distribution in conversations. Though Twitter has a hash-tag based mechanism to identify the topics of a tweet, it is impractical to keep track of huge number of hash-tags. Further, due to the large variations in hash-tags, the distribution will be sparse. Thus, we use Latent Dirichlet Allocation (LDA) [25] which is widely used for the topic modeling, to extract the hidden topics from the textual data. LDA is a Bayesian probabilistic model of text documents. It assumes a collection of $'k'$ topics and each topic defines a multinomial distribution over the vocabulary. To obtain a generic topic model, we perform topic modeling using all the tweets in the corpus. For language models, perplexity is used as a measurement of how well a probabilistic model predicts the sample. A lower perplexity score indicates better generalization performance of the probabilistic model. In order to tune the hyperparameter $'k'$, we conducted experiments by varying values of $'k'$ from 4 (as we already have 4 broad topic categories, we want to extract at least these number of latent topics) to 50 and calculated the perplexity of the resulting models. While performing these experiments on the dataset as we increased the value of $'k'$, perplexity score started to flatten out after $k = 15$. Thus, we used $k = 15$ for our experiments. An important point to note here is that the value of $'k'$ will vary with the dataset. In a much larger dataset, $'k'$ will be higher as there will be more hidden topics as compared to our dataset.

To calculate the topic distribution of the message content, we apply the topic model obtained using the aforesaid process to the given message content. This gives us the following feature:

- k-dimensional real-valued vector representing the topic distribution of the message $'m'$.

3.2.2. Latent User Preferences (LUP)

Most of the users post tweets and retweets on variety of topics of their interest [26,27]. With the aim of modeling the interests of a user regarding these topics, we apply the LDA based topic model (Section 3.2.1) to all her past tweets and retweets and calculate the topic distribution of her tweets. The resulting topic vector reflects latent user interests and preferences. Such topic vector is computed for both of the users involved in the link formation, i.e., the source user and the target user. Thus, we obtain the following features:

- k-dimensional real-valued vector representing the interest distribution of the source user $'s'$,
- k-dimensional real-valued vector representing the interest distribution of the target user $'d'$.

3.2.3. Network Features (NF)

To study the information diffusion, apart from the features based on the content of tweets, it is necessary to consider the features providing information about social network dynamics. The social network dynamics related to the users involved in the link is represented using the following features.

*Connections:*The study by Suh et al. [28] shows that the features related to social network like number of friends, number of followers, and number of user mentions are good indicators of 'retweetability'. Due to this reason, we include the following features:

- number of followers, and number of friends of the source user $'s'$,
- number of followers, and number of friends of the target user $'d'$.

Activity and response: To capture the social influence among users in the network, we include the following properties:

- total number of tweets posted by the source user $'s'$, to capture her activity,
- total number of retweets of tweets posted by the source user $'s'$, to capture the response she gets from other users in the network,
- total number of retweets made by the target user $'d'$, to capture her responsiveness towards other users.

In order to account for the topic sensitive information for the social influence, we include the following:

- total number of tweets made by the source user $'s'$ related to the topic $'t'$,
- total number of retweets of tweets posted by source user $'s'$ which are related to the topic $'t'$,
- total number of times target user $'d'$ has retweeted the tweets related to topic $'t'$.

Interaction: To capture the interaction between the source user and the target user, we include the following as a feature:

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

- number of times target user $'d'$ has reweeted the tweets of source user $'s'$.

Thus, we obtain 11 features representing the social network dynamics related to the users involved in the link.

3.2.4. Similarity between Users (SU)

Fei et al. [13] have shown that content propagation across links occurs more frequently between the users having common interests and thus, the content dissemination process is driven by the interest based homophily in the social network [19]. For computing this feature, we use the latent user preferences (Section 3.2.2), to represent the interest distribution of a user. To find the interest similarity between two users, we calculate the cosine distance between the interest distributions of the users. Thus, the similarity between users can be defined as:

$$S_{s,d} = \frac{I_s.I_d}{\|I_s\|\|I_d\|} \tag{1}$$

Where, I_s, I_d represent the interest distribution as obtained using topic model (Section 3.2.2) for user s and user d, respectively. This gives us the following feature:

- a real-valued feature representing the interest similarity of the source user $'s'$ and the target user $'d'$.

3.2.5. Content User Similarity (CUS)

Romero et al. [20] have demonstrated that the probability of information diffusion through a node increases with increase in the similarity between the textual content of the message under consideration and the user interests. For computing this feature, we use the latent user preferences (Section 3.2.2), to represent the interest distribution of a user, and latent topic information (Section 3.2.1), to represent the topic distribution of the message content. To find the similarity between the message content and the user interests, we calculate the cosine distance between the topic distribution of the message content and the interest distribution of the user. Thus, the content user similarity can be defined as:

$$S_{m,u} = \frac{T_m.I_u}{\|T_m\|\|I_u\|} \tag{2}$$

Here, T_m, I_u represent the topic distribution of the message content and the interest distribution of the user, respectively. Using this approach, we calculate the following two features:

- a real-valued feature representing the interest similarity of the source user $'s'$ and the message $'m'$,
- a real-valued feature representing the interest similarity of the target user $'d'$ and the message $'m'$.

3.3. Modeling Bayesian network

In this section, we describe the Bayesian network based method to predict the probability of diffusion of a given message through the links in social network. Bayesian classifiers are probabilistic classifiers. These are used to predict the probabilities of class belongingness, i.e., the probability that a given input data belongs to a particular class. Some of the important concepts related to Bayesian network are explained below, these are useful in building the proposed model.

3.3.1. Related concepts

(i) **Naive Bayes classifier** is the simplest Bayesian classifier and it assumes that all the feature values are conditionally independent of one another. However, in real situations, the assumption of conditional independence does not hold for every pair of features, i.e., some dependencies can exist among features. In such cases, Bayesian belief networks can be used.

(ii) **Bayesian belief networks** are graphical models. These are different from the Naive Bayes in the sense that the dependencies among subsets of features are considered in these models. Thus, they represent joint conditional probability distribution of the features. Moreover, these allow certain subsets of variables to remain conditionally independent.

(iii) A **Bayesian network topology** is a directed acyclic graph. Each node in the Bayesian network corresponds to a random variable (feature) and each edge represents a probabilistic dependence between these features. The edges are directed from parents to descendants.

(iv) An important **Bayesian network property** is *"Each attribute is conditionally independent of its non-descendants in the network, given its parents [29]."*

(v) Bayes theorem, stated by (3), is the underlying basis of Bayesian classification model. **Bayes theorem:** The probability of data tuple **D** belonging to a specified class C is given by:

$$P(C|\mathbf{D}) = P(C)\frac{P(\mathbf{D}|C)}{P(\mathbf{D})} \tag{3}$$

(vi) **Modeling continuous-valued features:** A continuous-valued attribute is generally assumed to have a Gaussian distribution [29], with a mean μ and standard deviation σ, as defined below

$$g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{4}$$

Hence, for an attribute A_k, the probability that data point $A_k = x_k$ belongs to class C_i is computed as:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \tag{5}$$

where, μ_{C_i} and σ_{C_i} are the mean and standard deviation, respectively, calculated from the values of feature A_k in the data tuples belonging to class C_i.

3.3.2. Learning Bayesian network

In order to learn the Bayesian network, it is important to know the dependence among various features under consideration. For this purpose, we compute Pearson correlation coefficient (PCC) matrix using the training data. PCC measures the linear dependence between two variables X and Y, in the range $[-1, +1]$, where -1 is total negative correlation, 0 is no correlation, and $+1$ is total positive correlation. PCC can be computed using the following Eq. (6).

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{6}$$

where, $cov(x, y)$ is the covariance and σ_x, σ_y are the standard deviations of x, y, respectively.

The features having "good" correlation coefficient values are considered to be dependent. The direction of dependency is decided in terms of action-cause relationship that is based on the general understanding of human behavior. The cut-off value of the correlation coefficient that can be considered as "good" was chosen experimentally. We performed experiments by varying this cut-off value of correlation from 0.1 to 0.4 and learned the Bayesian network using the features having absolute value of correlation coefficient greater than the chosen threshold. We rank all the Bayesian networks thus obtained as per Bayesian information criterion (BIC) score [30] to find the best fit. From these experiments, we found that the cut-off value 0.2 gives a good-fitting Bayesian network with the lowest BIC score.

Using this approach, we obtained the Bayesian belief network, as shown in Fig. 2, and used it for building our solution model. Each node in the network corresponds to a feature and edges represent probabilistic dependence between these features. The features which do not appear in the network are assumed to be conditionally independent of others.

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

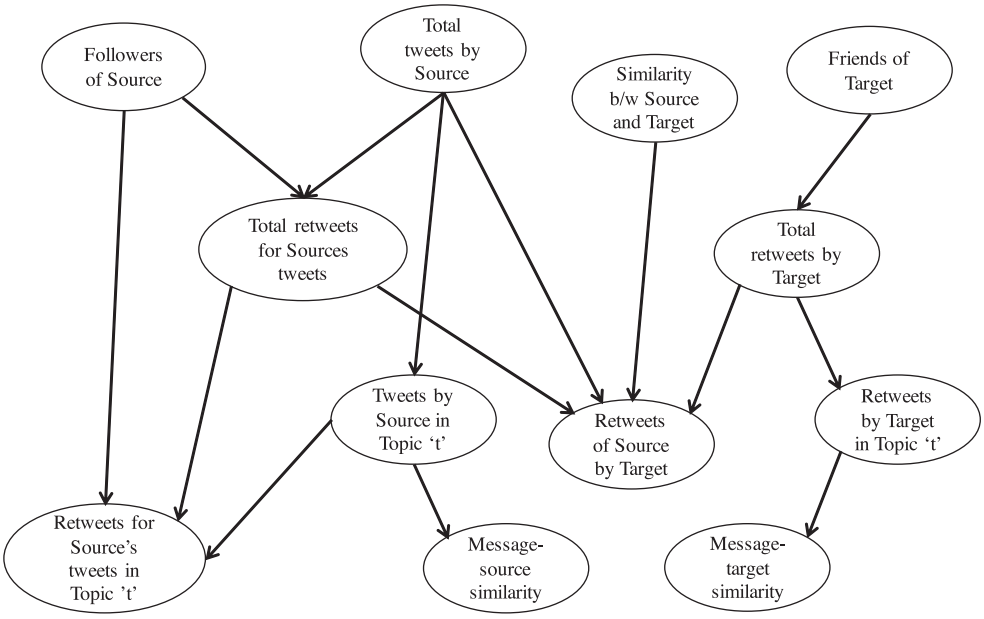

Fig. 2. Bayesian network topology used for the model.

After obtaining the Bayesian Belief Network, we employ the following approach to train the model. Suppose $(X) = (x_1, x_2, \ldots, x_n)$ is a data tuple entry for the feature-set $(A_1, A_2, \ldots, A_n)$. Using the Bayesian network property (Section 3.3.1 - iv), the joint distribution $P(x_1, x_2, \ldots, x_n)$ can be defined by (7) as follows:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(A_i)) \tag{7}$$

After applying the Bayes theorem (3) on the right hand side of (7), it can be rewritten in the following form (8):

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \left\{ P(x_i) * \prod_{j \epsilon Parents(A_i)} \frac{P(x_i, x_j)}{P(x_i) * P(x_j)} \right\} \tag{8}$$

where, $P(x_i, x_j)$ is the joint bivariate normal distribution, which is defined by (9) as below:

$$P(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} exp\left[-\frac{1}{2(1-\rho^2)} Q(x, y) \right] \tag{9}$$

$$where, Q(x, y) = \left(\frac{x - \mu_x}{\sigma_x} \right)^2 + \left(\frac{y - \mu_y}{\sigma_y} \right)^2 - 2\rho \left(\frac{x - \mu_x}{\sigma_x} \right) \left(\frac{y - \mu_y}{\sigma_y} \right) \tag{10}$$

here, μ_x and μ_y denote the means of x and y, respectively; σ_x and σ_y denote the standard deviations of x and y, respectively and ρ denotes the correlation coefficient between x and y.

The aim of a Bayesian classifier is to predict the probability that given data tuple **D** belongs to a specified class C. Hence, the Bayesian classifier can be built using the Bayes theorem (3). As $P(\mathbf{D})$ is constant for all classes, it is usually ignored. $P(C)$ is the class prior probabilities, it can be computed as support values using the training data. The value $P(\mathbf{D}|C)$ can be computed from the training data using the Equations from (5) to (10). Thus, in the process, labeled data can be used for calculating all the required parameters, namely, mean (μ), standard deviation (σ), and correlations (ρ) of different features for both the positive and negative classes. For our purpose, class $C = 1$ represents the active edges in the diffusion and $C = 0$ represents the inactive ones. **D** is the tuple of feature values as obtained in Section 3.2. Algorithm 1 presents the method to learn various parameters required for Bayesian classifier model.

3.3.3. Testing Bayesian network

After learning the required parameters, the probability of class belongingness for any given data tuple can be computed using (3) and (8). Algorithm 2 presents the steps for the same.

For our solution, we have used all the features described in Section 3.2, i.e., LTI, LUP, NF, SU and CUS. We have also used the dependencies among features as given in Fig. 2. Using this approach, model learns and predicts the probability of a given diffusion record (s, d, m, t) being active, i.e., given a message m related to topic t which is diffused by the user s, it predicts the probability that user d will further propagate (retweet) the message.

4. Experimental evaluation

4.1. Data preparation

After collecting the data (Section 3.1), several pre-processing steps are performed for data cleansing. The textual content obtained is processed by stemming the words and removing stopwords and symbols @ and RT which have specific usage in twitter messages. This processed data is then used for model formulation and experiments.

We collected data using various keywords related to different topic categories as discussed in Section 3.1. We run the experiments on different topic category datasets. The term *topic category dataset* represents the dataset where the diffusion records of the messages related to that topic category are taken as positive instances for model training and testing and others are considered negative. For example, in Social media category dataset, the following diffusion record is considered as positive instance.

(anon3, anon4, "Facebook buys WhatsApp for USD19 billion", Social Media)

while the following diffusion record is considered negative.

(anon3, anon4, "Thanksgiving is a time for family & fun", Festive Season)

While performing an experiment for a particular category, topic sensitive features related to activity and response (Section 3.2.3)

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

```
#data : matrix where each row (x1, x2, ..., xn) is a data tuple
#    entry for feature-set (A1, A2, ..., An)
#data_0 : list of data tuples with class label '0'
#data_1 : list of data tuples with class label '1'
#n : number of features
#P_i : priori probability of class i
#mu_i : array containing mean of feature values belonging to class i
#std_i : array containing standard deviation of feature
#   values belonging to class i
#corr_i : correlation coefficients matrix for feature
#   values belonging to class i
function learning_bayesian(data){
    mu = [0.0]*n
    std = [0.0]*n
    corr = [[0.0]*n]*n
    trans_data = transpose(data)
    for i in range(0,n):
        mu[i] = mean(trans_data[i])
        std[i] = sqrt(variance(trans_data[i]))
    for i in range(0,n):
        for j in range(i+1,n):
            corr[i][j] = correlation(trans_data[i],trans_data[j])
            corr[j][i] = corr[i][j]
    return mu,std,corr
}

#learn parameters for Class '0' model and Class '1' model
P_0 = len(data_0)/(len(data_0)+len(data_1))
mu_0, std_0, corr_0 = learning_bayesian(data_0)
P_1 = len(data_1)/(len(data_0)+len(data_1))
mu_1, std_1, corr_1 = learning_bayesian(data_1)
```

Algorithm 1. Learning parameters for Bayesian network model.

are calculated as per data restricted to that topic, while calculation of other features remains the same for all topics.

4.2. Experiments

Our model predicts the probabilities of information diffusion through various links in the social network, and true values of these probabilities are not available in the real world datasets. Thus, model evaluation based on relative error calculation is not feasible. Moreover, it is more important to predict accurate diffusion cascades rather than obtaining exact values of diffusion probabilities. Thus, we employ alternative evaluation methods based on the true values available in the dataset, i.e., the active links during a diffusion. We perform two types of experiments to evaluate our model, as described below.

Experiment 1 (Classification based): We perform classification based experiment to evaluate the performance of the features considered in building the proposed model. This experiment classifies the links in the network as active or inactive for the diffusion of the message under consideration based on the probability computed by the proposed model. Thus, if probability ≥ 0.5 then that link is labeled as active, otherwise inactive. Multiple such experiments are performed to evaluate the performance of the features on different topic category datasets. This addresses the problem similar to the one presented by Kuo et al. [14].

Evaluation Metrics: To better evaluate the performance of the classifier, we use $F1$ -score and the area under the curve (AUC) of a receiver operating characteristic (ROC) curve as the measures of accuracy [31]. The $F1$ -score is harmonic mean of precision and recall and is a standard metric of performance evaluation in the information retrieval field. The ROC curve has "True positive rate (sensitivity)" along y-axis and "False positive rate" along x-axis. So, the larger the values of $F1$ -score and area under the ROC curve, the better the classifier performance. We use a 10-fold cross validation

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

```
#test_data : data tuple for which probability is to be computed
#parents : Bayesian network in adjacency list representation
#proba_i : probability that given data tuple belongs to class i
#   as per model of class i
#result : normalized probability of data tuple to belong in class 1
function bayesian_model(data, mu, std, corr){
    p=1.0
    for i in range (0,n):
        p_i = normal_pdf(data[i], mu[i], std[i])
        p = p*p_i
        if parents[i]:
             for each j in parents[i]:
                p_j = normal_pdf(data[j], mu[j], std[j])
                p_joint = joint_pdf(data[i], data[j], mu[i],
                    mu[j], std[i], std[j], corr[i][j])
                p_conditional = p_joint/(p_i * p_j)
                p = p*p_conditional
    return p
}
#Testing with Class '0' model
proba_0 = P_0 * bayesian_model(test_data, mu_0, std_0, corr_0)
#Testing with Class '1' model
proba_1 = P_1 * bayesian_model(test_data, mu_1, std_1, corr_1)
#normalize the probabilities
#result is probability of test_data to belong in class '1'
result = proba_1/(proba_1+proba_0)
```

Algorithm 2. Testing Bayesian network model to compute IDP.

process for performance evaluation on each topic category dataset to remove any bias of over-fitting.

Baseline: We compare the performance of the proposed model with the model proposed by Kuo et al. [14]. For the comparative analysis, we consider the best performing feature set that includes topic similarity (TS), indegree (ID), and number of distinct topic propagated through the link (NDT) of the model presented by Kuo et al. [14]. To the best of our knowledge, this is one of the recent papers which performs classification of social network links as active or inactive for information diffusion using both content and network based features, thus, it is considered as the baseline for these experiments. This model is applied to all the instances present in our experimental dataset for different topic categories.

Results: Table 4 summarizes the results of classification based experiments for different topic category datasets. The results using the proposed Bayesian network model as well as using the baseline model [14] are shown. The features used in the proposed model are chosen based on the results of analysis performed by Varshney et al. [24]. Varshney et al. [24] addressed the classification of social network links as active or inactive. They analyzed the performance of various combinations of feature for different categories of datasets using support vector machines and artificial neural network based classifiers and verified that these features are effective. Thus, features used in the proposed model are more effective in capturing the information required to analyze the message diffusion through the links in a social network. Additionally, the model outperforms the baseline model by significant margins. Thus, we claim that the probability computed using proposed method can reliably classify the active/ inactive links for information diffusion.

Table 4
Diffusion probability: classification results.

	Kuo et al. [14] Baseline		Proposed model	
Topic ategory	F1-score	AUC(ROC)	F1-score	AUC(ROC)
Shopping	52.71%	74.49%	78.44%	90.58%
Politics	53.92%	76.80%	74.20%	85.97%
Social media	48.76%	72.04%	63.85%	83.21%
Festive season	50.88%	72.24%	63.84%	84.79%

Experiment 2 (Spread prediction based): The use of deterministic classification based experiment is insufficient for evaluating a probabilistic model. In order to perform probabilistic evaluation of the proposed Bayesian network based model, we perform experiments to predict the diffusion spread according to Independent

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

```
#seed_users : initial set of users who are active
#target : set of users who are active but yet to diffuse the message to their
    neighbors
#processed : set of users who have been processed for diffusion
#original_labels : original active/inactive labels for various links in the
    network

#Step 1: Find top 100 influential nodes in the network for each topic category
    based on the number of times their tweets got retweeted.
seed_users = find_top_100_influential_nodes(topic)

#Step 2: Perform the following steps to simulate the information spread 1000
    times and compute the average accuracy scores.
for i = 1 to 1000:
    function simulate_spread(seed_users){
        target = seed_users
        processed = NULL
        for each u in target:
            target.remove(u)
            processed.push(u)
            for each v in neighbors(u):
                r = random(0,1)
                p = diffusion_probability(u,v)
                if p>0 and p>=r:
                    predicted_label[u,v] = 1
                    if v not in processed:
                        target.push(v)
        return compute_accuracy_metrics(predicted_labels, original_labels)
    }
```

Algorithm 3. Monte Carlo simulation based experiment.

Cascade Model (ICM) of information diffusion. In this experiment, 1000 simulations are performed using Monte Carlo method [32]. Monte Carlo procedure simulates drawing values for the model parameters, which is activation threshold in ICM. By performing a large number of Monte Carlo simulations, it automatically accounts for the uncertainty that can arise due to external factors. The experiment is summarized by Algorithm 3:

Evaluation metrics: To better evaluate the performance of the classifier, we use Precision, Recall and $F1$ -score as the measures of accuracy [31].

Baseline: We compare the performance of proposed model with the approach proposed by Jiang et al. [17]. To the best of our knowledge, this is the most recent paper which predicts the retweeting probability score, thus, it is considered as the baseline for these experiments. For comparison, we run these experiments once by calculating the diffusion probabilities of links present in the network as per their approach and once by using our approach. Other works with matching objectives are [8,16]. But, these methods do not take content of the post into consideration. Also, the work in [8] is based on some assumptions such as a user performs an action at most once and the influence graph is DAG, which makes this method less suitable for tweet-retweet networks. Further, the methodology presented in [16] is not scalable to huge datasets, because EM algorithm used by them has to update the diffusion probability of each link in each iteration. In the following sections, we have presented the performance comparison with [17].

Results: Table 5 summarizes the results of information spread simulation based experiments for different topic category datasets. The results clearly indicate that proposed model outperforms the baseline method by significant margins. These results also indicate that our features significantly capture various topic related aspects of diffusion in addition to user preferences and network dynamics. Thus, we claim that these probabilities can be provided as input to influence maximization algorithms for obtaining improved results.

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003

Table 5
Diffusion probability: information spread simulation results.

	Jiang et al. [17] Baseline			Proposed model		
Topic ategory	Precision	Recall	F1-score	Precision	Recall	F1-score
Shopping	81.80%	39.44%	53.21%	86.36%	83.44%	84.87%
Politics	90.06%	37.82%	53.24%	78.34%	91.54%	84.43%
Social media	82.97%	36.27%	50.47%	84.79%	82.00%	83.37%
Festive season	79.72%	39.72%	53.02%	78.08%	90.81%	83.96%

4.3. Running time:

The experiments were performed on a machine having 64-bit Xeon processor with 4 physical cores and 16 GB memory. The environment was Windows 7 and Python 2.7. The most crucial and time consuming part of this solution framework is the step of learning LDA based topic model using all the collected tweets and retweets as discussed in Section 3.1. For this purpose, we used a fast implementation of online LDA [33] provided by the *gensim*[4] module. It took us approximately 4 h 6 min to compute the LDA model. We can further speed up this step by using multi-core version of LDA implementation[5] available in *gensim* module, which uses multiple cores to parallelize the training process. After computing LDA model, calculation of the feature vectors for all the instances in corpus requires only simple arithmetic calculations or database lookups, which took us 3435 s in total. After obtaining the feature vectors, we can learn the Bayesian model using the Algorithm 1. The learning algorithm involves statistical computations over the feature vectors and it scales linearly with the number of diffusion records in the training data. Same is the case with testing procedure (Algorithm 2). Overall, Experiment 1 took 3760 s to run with 10-fold cross validation for all 4 topic categories, while Experiment 2 on an average took around 15 h to run 1000 simulations for each topic category.

5. Discussion

While information diffusion in social networks is a widely studied topic in both computer science and sociology, the methods to quantify the diffusion probabilities of various social ties are not yet established. In this paper, we studied the information diffusion process and carried out experimental studies using Twitter data, to answer the research questions stated in Introduction section:

- **RQ1**: *What factors related to network settings, user characteristics, information source and message content affect the IDP values? How one can extract information out of the data available from online social networks to quantify these factors as features?*
  We studied the related literature and investigated various factors that play significant role in shaping the diffusion of a message through social network. A list of such factors is presented in Section 3. We have also discussed about the importance of the features related to network topography, user interests, content of the message and source of information in studying the information diffusion process in Section 3. Although, not all data is publicly available in case of online social networks. Most of the datasets are proprietary to some organizations. In this work, we presented approaches to compute different features from the limited data which is publicly available. The methods to extract out as much information as possible from collected dataset and approach to compute the feature values are discussed in Section 3.2.
- **RQ2**: *Whether these features are independent or there exist some dependencies among these and how one can formulate these dependencies?*
  We have performed statistical analysis to identify correlation among these features. Using Pearson Correlation Coefficient and Bayesian belief networks, we have empirically presented the significant dependencies existing among these features in Section 3.3.2. We found that while some of the features are independent, others have dependencies among them.
- **RQ3**: *How to design a model which can leverage these features and dependencies to reliably predict the IDP values in a social network for a given message?*
  Once the features and dependencies among them are obtained, we have evaluated various approaches to leverage this information in order to train a robust model for predicting IDPs. From experimental studies, we found that Bayesian classifier is best fit for this purpose since we needed a probabilistic classifier and we had to consider the dependencies among various features. In Section 3.3, we have presented the details to design the Bayesian network based model for this task. Further, results of different experiments conducted on real dataset established the reliability of this method.

We demonstrated the effectiveness of our solution with extensive experiments on a Twitter (sub-)network dataset. From experimental results (Section 4.2), we can draw following observations (1) Performance of the proposed method is consistently good across all the datasets for all topic categories under consideration revealing its effectiveness in capturing factors deciding information diffusion, (2) it can achieve better performance than state-of-the-art methods, which indicates that a method needs to consider the content of the information diffused and user interests along with the features related to network dynamics in order to reliably predict the information diffusion. Therefore, we propose that the diffusion probabilities obtained using our model can be provided as inputs to the applications like influence maximization, recommendation systems, trust propagation, etc. for a better performance. Moreover, running times for experiments (Section 4.3) indicate that the proposed method is efficient enough to be applied on a large dataset.

Although we have presented experiments on Twitter data only, this model can also be successfully applied to the data collected from other social networks and blogs. In the case of Facebook, like retweet action in Twitter, we can consider multiple actions (like, share, react, comment) as diffusion records, for studying the information diffusion. One might need to assign different weights to these actions in order to evaluate the strength of diffusion link. In the case of blogs, a successful diffusion can be considered when a user comments on the blog of another user or puts a link in her article pointing to some other blog or includes some words/ phrases from other user's blog post.

However, our approach is not without limitations. The proposed model does not perform well for users who are new in the social network or links which do not have any prior diffusion records. Also, this approach does not take the time elapsed into account,

[4] http://radimrehurek.com/gensim/ .
[5] https://radimrehurek.com/gensim/models/ldamulticore.html .

while we believe taking time delay into account will further improve the performance.

6. Conclusion

In this paper, we addressed the problem of inferring information diffusion probabilities in a social network. The contributions of this work are manifold. Firstly, we argued that the information diffusion probabilities cannot be accurately predicted using the network dynamics alone, this task needs to consider message contents and user interests also. Then, we presented a feature extraction process which utilizes latent topic information present in the content of tweets. Next, we found the dependencies that exist among various features to model Bayesian network. We validated the proposed Bayesian learning model on real world dataset obtained from Twitter. We conducted multiple experiments on different topic category datasets to evaluate our approach. Results of these experiments indicate the effectiveness of the proposed approach to compute the spread of information in the social network.

In future, we would like to address the limitations of our approach for predicting the diffusion probabilities of newly formed links which do not have any prior diffusion records. In addition, we plan to model the cascade and the delay in diffusion of information in the social network. Efforts will be done in finding the time by which a user is expected to perform an action to propagate the information further.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the anonymous reviewers for their valuable comments that helped improve the quality of this paper.

References

[1] A.L. Hughes, L. Palen, Twitter adoption and use in mass convergence and emergency events, Int. J. Emergency Manage. 6 (3–4) (2009) 248–260.
[2] P.N. Howard, A. Duffy, D. Freelon, M.M. Hussain, W. Mari, M. Maziad, Opening closed regimes: what was the role of social media during the arab spring? Project Inf. Technol. Polit. Islam (2011) 1–30.
[3] W. Chen, W. Lu, N. Zhang, Time-critical influence maximization in social networks with time-delayed diffusion process., in: AAAI, vol. 2012, 2012, pp. 1–5.
[4] M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi, The world of connections and information flow in twitter, IEEE Trans. Syst. Man Cybern. Part A 42 (4) (2012) 991–998.
[5] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 57–66.
[6] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information diffusion in online social networks: a survey, ACM SIGMOD Record 42 (2) (2013) 17–28.
[7] K. Zhang, X. Yun, J. Liang, X.-y. Zhang, C. Li, B. Tian, Retweeting behavior prediction using probabilistic matrix factorization, in: Computers and Communication (ISCC), 2016 IEEE Symposium on, IEEE, 2016, pp. 1185–1192.
[8] A. Goyal, F. Bonchi, L.V. Lakshmanan, Learning influence probabilities in social networks, in: Proceedings of the third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 241–250.
[9] M. Taherian, M. Amini, R. Jalili, Trust inference in web-based social networks using resistive networks, in: Internet and Web Applications and Services, 2008. ICIW'08. Third International Conference on, IEEE, 2008, pp. 233–238.
[10] J.J. Samper, P.A. Castillo, L. Araujo, J. Merelo, Nectarss, an rss feed ranking system that implicitly learns user preferences, Comput. Res. Reposit. Vol. abs/cs/0610019 (2006).
[11] C. Kang, C. Molinaro, S. Kraus, Y. Shavitt, V. Subrahmanian, Diffusion centrality in social networks, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 558–564.
[12] D. Kempe, J. Kleinberg, É. Tardos, 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, 2003, pp. 137–146.
[13] H. Fei, R. Jiang, Y. Yang, B. Luo, J. Huan, Content based social behavior prediction: a multi-task learning approach, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 995–1000.
[14] T.-T. Kuo, S.-C. Hung, W.-S. Lin, N. Peng, S.-D. Lin, W.-F. Lin, Exploiting latent information to predict diffusions of novel topics on social networks, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 344–348.
[15] J. Zhu, F. Xiong, D. Piao, Y. Liu, Y. Zhang, Statistically modeling the effectiveness of disaster information in social media, in: Global Humanitarian Technology Conference (GHTC), 2011 IEEE, IEEE, 2011, pp. 431–436.
[16] K. Saito, R. Nakano, M. Kimura, Prediction of information diffusion probabilities for independent cascade model, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2008, pp. 67–75.
[17] B. Jiang, J. Liang, Y. Sha, R. Li, W. Liu, H. Ma, L. Wang, Retweeting behavior prediction based on one-class collaborative filtering in social networks, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2016, pp. 977–980.
[18] S. Wen, J.J. Jiang, Y. Xiang, S. Yu, W. Zhou, Are the popular users always important for information dissemination in online social networks? IEEE Netw. 28 (5) (2014) 64–67.
[19] S.A. Golder, D.M. Wilkinson, B.A. Huberman, Rhythms of social interaction: Messaging within a massive online network, in: Communities and Technologies 2007, Springer, 2007, pp. 41–66.
[20] D.M. Romero, B. Meeder, J. Kleinberg, Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter, in: Proceedings of the 20th International Conference on World Wide Web, ACM, 2011, pp. 695–704.
[21] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 61–70.
[22] C. Lin, Q. Mei, Y. Jiang, J. Han, S. Qi, Inferring the diffusion and evolution of topics in social communities, Soc. Netw. Min. Anal. 3 (d4) (2011) d5.
[23] S.-C. Hung, T.-T. Kuo, S.-D. Lin, Novel topic diffusion prediction using latent semantic and user behavior, in: Proceedings of the ASE BigData & SocialInformatics 2015, ACM, 2015, p. 39.
[24] D. Varshney, S. Kumar, V. Gupta, Modeling information diffusion in social networks using latent topic information, in: Intelligent Computing Theory, Springer, 2014, pp. 137–148.
[25] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[26] N. Banerjee, D. Chakraborty, K. Dasgupta, S. Mittal, A. Joshi, S. Nagar, A. Rai, S. Madan, User interests in social media sites: an exploration with micro-blogs, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 1823–1826.
[27] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, 2007, pp. 56–65.
[28] B. Suh, L. Hong, P. Pirolli, E.H. Chi, Want to be retweeted? large scale analytics on factors impacting retweet in twitter network, in: Social Computing (socialcom), 2010 IEEE Second International Conference on, IEEE, 2010, pp. 177–184.
[29] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.
[30] G. Schwarz, et al., Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[31] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240.
[32] S. Raychaudhuri, Introduction to monte carlo simulation, in: Simulation Conference, 2008. WSC 2008. Winter, IEEE, 2008, pp. 91–100.
[33] M. Hoffman, F.R. Bach, D.M. Blei, Online learning for latent dirichlet allocation, in: Advances in Neural Information Processing Systems, 2010, pp. 856–864.

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003