Dynamic Analysis of User-Role and Topic-Influence for Topic Propagation in Social Networks

JING WANG^{ID} AND HUI ZHAO^{ID}

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Corresponding author: Hui Zhao (hzhao@mail.xidian.edu.cn)

ABSTRACT Hot events spread quickly on social networks. Predicting event diffusion on social networks, also known as topic propagation, is an important task. The two important factors that affect topic propagation are users and topics, and both users' roles and topics' influences are time dependent on social networks. However, existing studies have largely overlooked this fact, so topic propagation prediction is still a major challenge. In this paper, a Topic Propagation Prediction method is proposed based on Dynamic Analysis of user-role and topic-influence, named TPP-DA, which predicts the topic propagation on social networks from both users' and topics' perspectives. First, we introduce a temporal perspective to improve the static analysis to the dynamic analysis of user-role, which is more adaptable to the changeable user-roles on social networks. Second, we introduce a metric called the topic heat to dynamically analyze the topic-influence on a single user and social group. Third, we combine the dynamic analysis of user-role and topic-influence with a weighted probability model to accurately predict topic propagation trends. The weights are determined by the dynamic analysis of user-role and topic-influence. Finally, several experiments are conducted to evaluate TPP-DA. Compared with TPP, the average error rate of TPP-DA is reduced by approximately 33%, which proves the efficiency of TPP-DA.

INDEX TERMS Topic propagation, user-role, topic-influence, probability model, social networks.

I. INTRODUCTION

Today, social networks are playing an increasingly important role in people's lives. Hot events or topics spread rapidly in social networks. Analyzing hot events in social networks can help governments and companies in mining useful information or studying rumor diffusion models [1]–[3]. For example, for a good topic in social networks, e.g., ''Olympic Games'', we must predict how many people will pay attention to the topic in the future. In contrast, we should understand the propagation law of a rumor so that the government can refute it. Therefore, topic propagation prediction, which predicts the spread of topics, is an important task.

Topic propagation refers to how information is propagated through social networks [4]. It may be affected by the users and topics in social networks. Additionally, topic propagation can be affected by other factors, such as the community

The associate editor coordinating the review of this manuscript and approving it for publication was Vijay Mago^{ID}.

structure and the temporal information of social networks [5]–[7]. While user relationships are important parts of community structure, and the time-varying user interest and topic heat embody the temporal information of social networks, they also depend on users and topics. Therefore, we only focus on analyzing the impact of users and topics on topic propagation in this paper.

First, users have different interests in different topics in social networks. A user may have different roles in different topics, so he or she may have varying degrees of influence on other users regarding various topics. Meanwhile, users have many friends with the same or similar interests in social networks, and they interact with each other in various ways. They may also have different roles in their relationships and have varying degrees of influence on each other to affect topic propagation. Furthermore, the analysis of user-role should consider the changeability of users' interests or hobbies over time. In brief, user-role analysis should be considered in topic propagation prediction.

Second, the topic itself affects topic propagation. Different topics in social networks have different influences. These topics compete for users' attention. For example, many people focus on hot events; therefore, hot topics can spread quickly. However, other topics may diffuse slowly in social networks. Meanwhile, the heat of the topics will change over time. As the topic heat changes, the topic-influence also changes. Therefore, topic-influence should be considered when studying topic propagation.

There are some researches on topic propagation models, such as the linear threshold (LT) model [8], [9] and the independent cascade (LC) model [10]–[13]. Some works utilized epidemic models [14]–[16], data-driven models [17]–[20], various dynamical models [21]–[25] or deep learning neural networks [26]–[29] for topic propagation to explore information diffusion. However, these works did not conduct both user-role and topic-influence analyses, so they could be further improved. In our previous work [30], we proposed a user-role-based topic propagation prediction (TPP) model, but it analyzed user-role statically, and it did not consider topic-influence. Social networks are time-varying, and user-roles and topic-influences are time dependent. Existing studies on topic propagation have overlooked the time dependency of user-roles and topic-influences; predicting topic propagation trends remains a major challenge.

In this paper, we propose a Topic Propagation Prediction method based on the Dynamic Analysis of user-role and topic-influence, named TPP-DA. Its purpose is to predict the topic propagation in social networks considering both users' and topics' perspectives.

First, we analyze the inadaptability of our previous static user-role analysis, which is the user-roles change over time in social networks. To address the problem, a temporal perspective s introduced to static analysis. Then, user-role is analyzed dynamically. Similar to static user-role analysis, four user-role factors are incorporated to characterize user attributes along two dimensions. In one dimension, the user expert-factor and leader-factor are analyzed based on a single user behavior. In the other dimension, the user social-factor and similarity-factor are described based on social behavior.

Second, topics are time dependent, and they compete for users' attention. The topic itself also plays an important role in topic propagation. Topic-influence should be considered when predicting topic propagation. Topic-influence analysis consists of two parts. The first part is the topic-influence on a single user, and the second part is the topic-influence on a social group. We also introduce a metric called topic heat to calculate the topic-influences on a single user and social group.

Third, the dynamic analysis of user-role and topic-influence is combined with a weighted probability model to predict topic propagation trends more accurately. Here, the weights are determined by the dynamic analysis of user-role and topic-influence. Behavior probability, relationship probability and time probability are utilized to form the probability model, and they are calculated based on user behaviors, group relationships and time spans.

Finally, we present the algorithms in detail and analyze the computational complexities of the algorithms. Some experiments are conducted to evaluate TPP-DA. Compared with TPP (our previous topic propagation with static user-role analysis), the average error rate of TPP-DA is reduced by approximately 33%. The experimental results show its efficiency. Our main contributions have particular importance for topic propagation prediction. To summarize, the contributions of this paper are as follows:

1) The dynamic analysis of user-role and topic-influence is utilized to conduct topic propagation prediction. To the best of our knowledge, this paper is the first to predict topic propagation based on the dynamic analysis of user-role and topic-influence in social networks from both users' and topics' perspectives.

2) A temporal perspective is introduced to improve our previous static analysis to dynamically analyze on user-role, which is more adaptable to user-roles changing over time in social networks. Topic-influences are also dynamically analyzed on both single user and social group, which are calculated based on topic heat.

3) The dynamic analysis of user-role and topic-influence is combined with a weighted probability model to predict topic propagation trends more accurately. Some experiments are conducted to evaluate TPP-DA, and the experimental results show its efficiency.

The remainder of this paper is organized as follows. Section II describes the related works. Section III presents the topic propagation problem formulation. Section IV shows the dynamic analysis of user-role and topic-influence in detail. Section V gives the weighted probability model and topic propagation (TPP-DA) algorithm. Section VI evaluates the method. Finally, the conclusion is presented in Section VII.

II. RELATED WORKS

One of the main issues in social networks is to predict information diffusion. With the development of networking and informatization in recent years in particular, the study of topic propagation models has been a popular topic in academia.

Several representative topic propagation models are available, such as the linear threshold (LT) model [8], [9] and the independent cascade (LC) model [10]–[13]. For strong communities that can facilitate global diffusion by enhancing local intracommunity spreading, Nematzadeh *et al.* [8] investigated topic propagation with the linear threshold model. Guille and Hacid [9] proposed a practical solution that aimed to predict the temporal dynamics of diffusion in social networks. The approach was based on machine learning techniques and the inference of time-dependent diffusion probabilities from a multi-dimensional analysis of individual behaviors. Gray *et al.* [10] studied the effect of graph structure on the flow of information over a network using Watts's simple model of global cascades. Wang *et al.* [11] developed a user representation learning model to solve the

information diffusion prediction problem on social media. The model learned the role-based representations based on a cascade modeling objective and employed the matrix factorization objective of reconstructing structural proximities to regularize the representations. A diffusion model based on a cascade model framework was proposed to generate the retweeting network in [12]. Gao *et al.* [13] proposed a novel information-dependent embedding-based diffusion prediction (IEDP) model. The proposed model further learned the propagation probability of information in the cascade as a function of the relative positions of information-specific user embedding in the information-dependent subspace.

Due to the similarity between topic propagation in social networks and virus propagation on biological networks, epidemic models [14]–[16] have been widely applied to explore the information dynamics in social networks. Considering that the numerical solutions (both continuous and discrete) of traditional SIR-based (Susceptible-Infective-Removed) models cannot match the corresponding simulation results of node behavior, Rui *et al.* [14] extended the classic SIR model to the new SPIR (Susceptible-Potential-Infective-Removed) model by introducing the new concept of the potential spread set. In the diffusion process, a new proposed model characterized a practical reinfection-reemergence scenario caused by the change in social attributes for individuals [15]. Kong *et al.* [16] introduced a connection between generalized stochastic SIR models and self-exciting point processes in a finite population. However, nearly all of the previous epidemic models failed to sufficiently consider the influence of social ties. Subsequently, some studies using data-driven models [17]–[20] for topic propagation have emerged. An analysis method of the influence of the potential edges of topic propagation was studied using a simple topic propagation model of the networks [17]. Molaei *et al.* [18] proposed a novel heterogeneous deep diffusion (HDD) approach in which functional heterogeneous structures of the network were learned by a continuous latent representation through traversing meta-paths. Li *et al.* [19] proposed a diffusion model based on multiple messages and a multiplex network space. Stai *et al.* [20] defined informed Twitter users as those who have produced/reproduced tweets with a specific hashtag to more precisely capture real information propagation on Twitter.

The above topic propagation models do not generally follow the principle of the conservation of matter. Therefore, the majority of studies on topic propagation use various dynamical models [21]–[25]. Based on the propagation dynamics, Liu *et al.* [21] proposed a nonlinear dynamic emergency topic propagation system and mathematical model for public events. Cao *et al.* [22] investigated topic propagation from an evolutionary game-theoretic perspective, and they derived the evolutionary dynamics and evolutionarily stable states (ESSs) of diffusion. Farajtabar *et al.* [23] proposed a temporal point process model, COEVOLVE, which efficiently simulated interleaved diffusion and network events and generated traces

obeying the common diffusion and network patterns observed in real-world networks. Hu *et al.* [24] proposed a hydro-dynamic topic propagation prediction model (hydro-IDP) that exploits a hydrodynamic model to describe the spreading process of information on online social networks. Saha *et al.* [25] proposed the Competing Recurrent Point Process (CRPP), a probabilistic deep learning machine that unifies the nonlinear generative dynamics of a collection of diffusion processes and interprocess competition - the two ingredients of visibility dynamics.

In addition, owing to the significant recent successes of deep learning in multiple domains, attempts have been made to predict information diffusion by developing neural network-based approaches [26]–[29]. Chen *et al.* [26] proposed a deep multitask learning-based information cascade model (DMTLIC), which explicitly modeled and predicted cascades through a multitask framework with a novelly designed shared-representation layer. Wang *et al.* [27] proposed a novel sequential neural network with structural attention to model topic propagation. The proposed model explores both the sequential nature of an information diffusion process and the structural characteristics of a user connection graph. Mishra [28] proposed a recurrent neural network model by modeling a social cascade with a marked Hawkes self-exciting point process and then learned a predictive layer on top for popularity prediction using a collection of cascade histories. Sankar *et al.* [29] present a novel variational auto-encoder framework (Inf-VAE), which utilized powerful graph neural network architectures to learn social variables to predict the set of all influenced users.

With the increase in the interaction frequency on hot topics among users in social networks, users and topics play increasingly important roles in topic propagation. Because the user-roles and topic-influences are time dependent in social networks, predictions of topic propagation trends should account for this changeability. However, the existing models either rely on the probabilistic modeling of information diffusion based on partially known network structures or discover the implicit structures of diffusion from users' behaviors without considering the dynamic analysis of user-role and topic-influence; therefore, they need to be further optimized and improved. In this paper, based on our previous work [30], we introduce the dynamic analysis of user-role and topic-influence to the topic propagation prediction model, which predicts topic propagation in social networks, considering both users' and topics' perspectives to predict topic propagation trends more accurately.

III. TOPIC PROPAGATION PROBLEM FORMULATION

In this paper, we define topic propagation as predicting the number of people who will spread a topic on a certain day in the social network. Specifically, given a social network and a topic, we should be able to predict how many people will pay attention to the topic in a few days. In this section, we will give the problem formulation of the topic propagation.
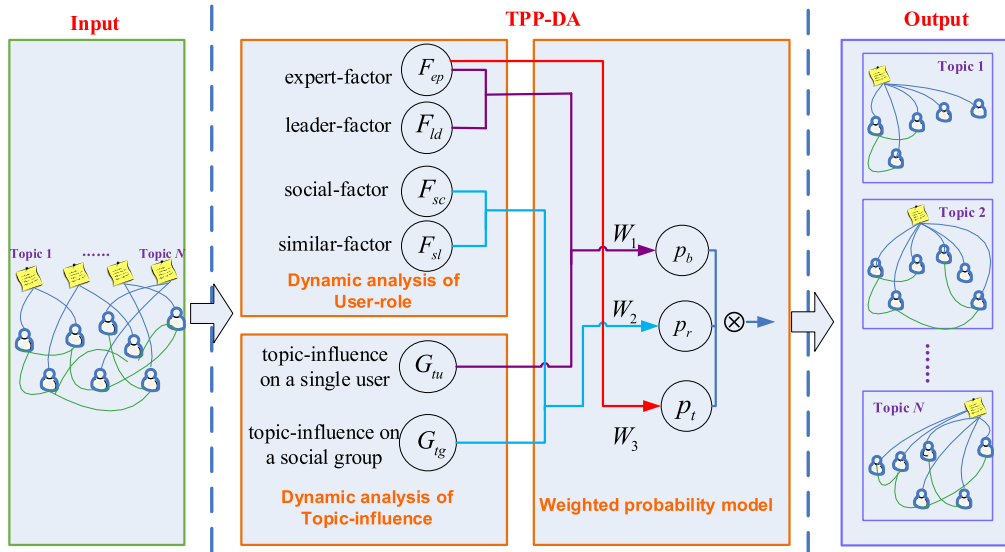
FIGURE 1. The topic propagation model framework.

We use $S = \{U, C, M\}$ to denote a social network. U represents all the users in the social network, $U = \{u_i \mid i = 1, 2, \ldots, N\}$ and u_i is the i-th user. C is the set of topics, and $C = \{z\}_{z=1}^{K}$. The topics can be automatically extracted with statistical topic modeling such as probabilistic latent semantic indexing (pLSI) [31] and latent Dirichlet allocation (LDA) [32]. M represents the set of all messages posted or commented on by all N users on all K topics, and $M = \{M_i^z \mid u_i \in U, z \in C\}$.

Since a social network is a time-varying system, let $U(t)$, $C(t)$ and $M(t)$ be the user set, topic set and message set at time t. $U(t) \in U$, $C(t) \in C$ and $M(t) \in M$. $M_i^z = \{m_i^z(t) \mid t = 1, 2, \ldots T\}$, which denotes the set of messages of user u_i on topic z from time 1 to time T, $m_i^z(t)$ is the set of messages of user u_i on topic z at time t, and $x_i^z(t) = \{0, 1\}$ denotes whether user u_i discusses topic z at time t. If $|m_i^z(t)| = 0$, then $x_i^z(t) = 0$; otherwise, $x_i^z(t) = 1$. For example, for user u_i and topic z, $\{x_i^z(1), x_i^z(2), \ldots, x_i^z(t), \ldots\} = \{0, 0, 1, 0, 1, \ldots\}$ means that user u_i discusses topic z at times 3 and 5. Another topic, z', $\{x_i^{z'}(1), x_i^{z'}(2), \ldots, x_i^{z'}(t), \ldots\} = \{1, 0, 0, 1, 0, \ldots\}$, signifies that user u_i discusses topic z' at times 1 and 4. For a topic z at any time t, if all of the users' $x_i^z(t)$ are summed, then we know how many users are discussing topic z at time t, i.e., $\sum_{i=1}^{N} x_i^z(t)$.

In brief, given a social network S with the users information (relationships) and all messages from time 1 to time T, and for a topic in S, we should predict how many people will pay attention to the topic at time t, $\sum_{i=1}^{N} x_i^z(t)$. Table 1 summarizes the important symbols used in the topic propagation model.

As mentioned earlier, the topic propagation may be affected by the users and topics in social network, and the user interest and topic heat change over time. Therefore, we introduce the dynamic analysis of user-role and topic-influence in topic propagation. Figure 1 shows the framework

TABLE 1. Important symbols used in the topic propagation model.

Symbol	Description
$S = \{U, C, M\}$	the social network
U	the user set in social network S
C	the topic set in social network S
M	the message set in social network S
u_i	a user in U
z	a topic in C
U_z	the set of users who have discussed topic z
M_i^z	the set of messages of user u_i on topic z
$m_i^z(t)$	the message set of user u_i on topic z at time t
$x_i^z(t)$	whether user u_i discusses topic z at time t

of our topic propagation model. The model contains three parts.

The first part is the dynamic analysis of user-role. Taking the social network as the input, user-roles are first dynamically analyzed from four aspects. Four user-factors are introduced to analyze user-role. The second part is the dynamic analysis of topic-influence, which takes the social network as the input and calculates the topic-influences on a single user and a social group. Based on the dynamic analysis of user-role and topic-influence, the third part introduces a weighted probability model to accurately predict topic propagation trends.

Based on the above three parts, a topic propagation model (TPP-DA) is established, which can predict how the topic will spread in social networks.

IV. DYNAMIC ANALYSIS OF USER-ROLE AND TOPIC-INFLUENCE

In this section, we introduce the dynamic analysis of user-role and topic-influence in topic propagation, which is the basis of the topic propagation model.

A. DYNAMIC USER-ROLE ANALYSIS

In our previous work [30], we studied how to statically analyze user-roles. However, since we know that social networks are time-varying systems and user-roles and topic-influences are time dependent in social networks too, the analysis of user-role should consider this changeability over time. For example, one of the user-roles, the expert-factor, has a value that has remained static and constant throughout time according to the static user-role analysis in our previous work. However, it would be impractical because users may have dynamic interests, attributes, and features over time. As a result, the expert-factor value will also change over time.

We calculate the expert-factor value using static user-role analysis with a dataset over 30 days. In addition, we divide the 30-day dataset into 10 consecutive sub-datasets, each spanning 3 days. Since we want to conduct static user-role analysis on topic propagation, we conduct the same calculation for the time windows (each 3-day sub-dataset) independently. That is, for each time window, we calculate the user expert-factors using data from every 3 days of the time window. These results are shown in Fig. 2. The static expert-factor over 30 days is constant, and the values every 3 days are different, which proves that user-roles change over time. This changeability should be considered when predicting topic propagation.
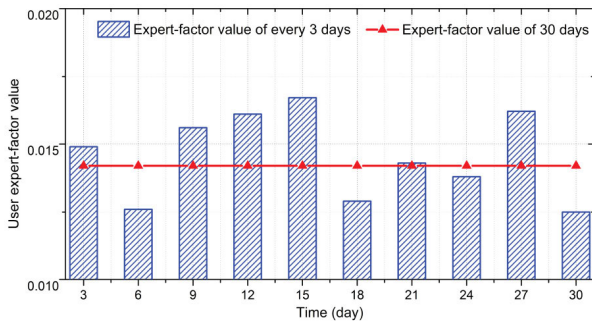
FIGURE 2. The static expert-factor values of 30 days and every 3 days.

In addition to the user expert-factor, the user leader-factor values are calculated over 30 days and every 3 days. The results are shown in Fig. 3. As the user leader-factor is calculated based on his/her relationships, which are generally stable, it is less influenced by time. As a result, the change in the user leader-factor is very small. In fact, the content-based factors (expert-factor and similarity-factor) are more time dependent, while the social-based factors (leader-factor and social-factor) are less time dependent. In summary, introducing a temporal perspective is necessary when analyzing user-role in social networks.

1) DYNAMIC USER EXPERT-FACTOR

This factor calculates the relative expertise of different users in a social network. The expert users in a social network exert a greater influence on topic propagation than other users. Based on our previous work, the dynamic expert-factor can
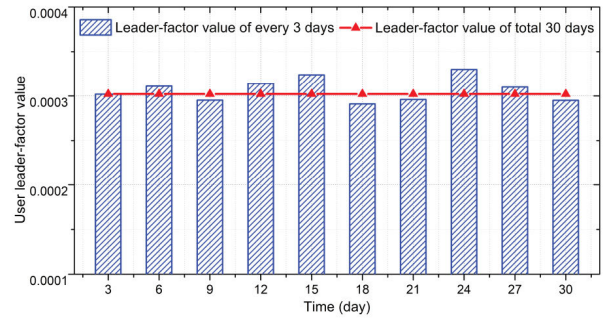
FIGURE 3. The static leader-factor values of 30 days and every 3 days.

be denoted by introducing the time dimension:

$$F_{ep}(u_i, z, t) = \frac{|m_i^z(t)|}{\sum_{u_j \in U(t)} |m_j^z(t)|} \qquad (1)$$

where $|m_i^z(t)|$ is the number of messages of user u_i discussing topic z at time t, and $\sum_{u_j \in U(t)} |m_j^z(t)|$ is the total number of messages of all users discussing topic z at time t.

2) DYNAMIC USER LEADER-FACTOR

This factor represents the influence of users' social relationships in social networks. Users with more social relationships may have greater influence than other users. Similarly, the static user leader-factor is modified by adding the time dimension. The PageRank algorithm is used to calculate all user leadership factors based on the regression relationship. According to the PageRank algorithm, the dynamic user leader-factor can be summarized as:

$$F_{ld}(u_i, z, t) = \frac{1-d}{|U_z(t)|} + d \sum_{j=1}^{k} \frac{F_{ld}(u_{i_j}, z, t)}{O(u_{i_j})} \qquad (2)$$

where $U_z(t)$ is the set of users who have discussed topic z at time t, $O(u_i)$ is the out-degree of u_i, and D is a dampening factor (generally $d = 0.85$). As with the iterative calculation of pagerank, all the user leader-factors are initially set to 1 and are then iteratively calculated according to the social relationships and Eq. (2) until they converge into stable values to get the final the values of user leader-factors. Clearly, $0 \leq F_{ld}(u_{i_j}, z, t) \leq 1$ and $\sum_i F_{ld}(u_{i_j}, z, t)$ for any topic z at time t.

3) DYNAMIC USER SOCIAL-FACTOR

The dynamic user social-factor measures the strength of all social relationships between the users in a social network. In other words, the more users who have the same friends, the closer and stronger the social relationships among users. Here, using time as the variable and adding parameter t, the dynamic user social-factor is obtained as:

$$F_{sc}(u_i, u_j, z, t) = \frac{fr(u_i, u_j, z, t)}{fr(u_i, u_j, t)} \qquad (3)$$

where $fr(u_i, u_j, z, t)$ is the number of mutual friends of u_i and u_j in $U_z(t)$ at time t, and $fr(u_i, u_j, t)$ is the number of mutual friends of u_i and u_j in $U(t)$ at time t.

4) DYNAMIC USER SIMILARITY-FACTOR

The dynamic user similarity-factor is defined as the number of messages that users interact with each other in social networks, which describes the similarity of each user pair in terms of subject preferences. The more interactions on the same topic, the more users exert similar influences in social networks. The dynamic user similarity-factor between u_i and u_j is calculated as:

$$F_{sl}(u_i, u_j, z, t) = \frac{|m_i^z(t)| + |m_j^z(t)|}{\sum_{z' \in C(t)} (|m_i^{z'}(t)| + |m_j^{z'}(t)|)} \quad (4)$$

B. DYNAMIC TOPIC-INFLUENCE ANALYSIS

In social networks, the topic cannot propagate without social users. In addition to users, the topic itself plays an important role in topic propagation. For example, hot topics can spread quickly while other topics may diffuse slowly in social networks. Generally, social networks have many topics. These topics have different influences, and they also compete for users' attention; therefore, topic-influence affects topic propagation. Some hot topics will spread to more users, which can further enhance their influences. As a result, the impact of topic-influence on topic propagation should be considered.

Topic-influence analysis consists of two parts. One part is the topic-influence on a single user, and the other part is the topic-influence on a social group. In this subsection, we analyze the topic-influence on both a single user and a social group.

1) TOPIC-INFLUENCE ON A SINGLE USER

In general, a user has dynamic interests; he or she may pay close attention to several topics that will consume the user's attention. Therefore, the first factor affecting the topic-influence on a single user depends on how much attention the user directs on the topic, i.e., the topic heat on the user. The following formula is adopted to calculate the topic heat on a user, which is the ratio of the user's messages on a certain topic to the user's messages on all topics at time t.

$$Heat(z, u_i, t) = \frac{|m_i^z(t)|}{\sum_{z' \in C} |m_i^{z'}(t)|} \quad (5)$$

Furthermore, users can interact with each other in social networks. If a user's friends are discussing a topic, then the user is highly likely to be interested in the topic; therefore, the topic-influence on a single user is related to the topic heat on his or her friends. As a result, the second factor affecting

the topic-influence on a single user is:

$$Heat(z, f(u_i), t) = \sum_{u_j \in f(u_i)} Heat(z, u_j, t)$$

$$= \sum_{u_j \in f(u_i)} \frac{|m_j^z(t)|}{\sum_{z' \in C(t)} |m_j^{z'}(t)|} \quad (6)$$

where $f(u_i)$ is the set of user u_i's friends.

Generally, the topic-influence of topic z on single user u_i at time t is:

$$G_{tu}(z, u_i, t) = Heat(z, u_i, t) \cdot Heat(z, f(u_i), t)$$

$$= \frac{|m_i^z(t)|}{\sum_{z'} |m_i^{z'}(t)|} \sum_{u_j \in f(u_i)} \frac{|m_j^z(t)|}{\sum_{z'} |m_j^{z'}(t)|} \quad (7)$$

2) TOPIC-INFLUENCE ON A SOCIAL GROUP

The topic-influence on social groups depends on the topic heat of topic z for all users in the social group. The topic heat on a social group is denoted as:

$$Heat(z, U, t) = \frac{\sum_{n=0}^{t} \sum_{u_i \in U} |m_i^z(n)|}{\sum_{n=0}^{t} \sum_{u_i \in U} \sum_{z' \in C} |m_i^{z'}(n)|} \quad (8)$$

If the total number of messages on topic z from time 1 to time t is large, then topic z may be popular, and its influence will also be large. Therefore, the topic-influence of topic z on a social group is:

$$G_{tg}(z, U, t) = Heat(z, U, t) \quad (9)$$

C. LAGRANGE INTERPOLATION POLYNOMIAL

For user u_i and topic z, if we have the previous s values, $(t, V(t)), (t-1, V(t-1)), \ldots, (t-s+1, V(t-s+1))$, then we can predict the next value $V(t+1)$. For example, if we have the previous s expert-factors $(t, F_{ep}(u_i, z, t))$, $(t-1, F_{ep}(u_i, z, t-1)), \ldots, (t-s+1, F_{ep}(u_i, z, t-s+1))$, then we can predict the value of $F_{ep}(u_i, z, t+1)$.

A Lagrange interpolation polynomial is used in this study. Given any finite set of points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with all x_i distinct, we can construct the unique polynomial of the minimum degree passing through all of those points. The Lagrange formula can be written as:

$$L(x) = \sum_{i=1}^{n} (y_i \prod_{j=1, j \neq i}^{n} \frac{x - x_j}{x_i - x_j}) \quad (10)$$

This formula works because each fraction takes the value of 1 at the appropriate x_i and takes the value of 0 at all other x_i.

Similarly, we can obtain a Lagrange interpolation formula $L_{ep}(u_i, z, x)$ given the previous s expert-factors, where time x is the variable in the Lagrange formula.

$$L_{ep}(u_i, z, x) = \sum_{k=0}^{s} [F_{ep}(u_i, z, t-k) \prod_{j=0, j \neq k}^{s} \frac{x - t + j}{(j - k)}] \quad (11)$$

Then, we can obtain the expert-factor of u_i on topic z at time $(t + 1)$, which is:

$$F_{ep}(u_i, z, t + 1) = L_{ep}(u_i, z, t + 1) \tag{12}$$

As for other user factors and topic-influences, we can construct similar Lagrange polynomials with previous s values and then obtain the dynamic value at time $t + 1$.

V. DYNAMIC TOPIC PROPAGATION MODEL

In this section, we present the weighted probability model used to predict topic propagation. First, we briefly introduce the probability model in our previous work [30]. Second, we study how to combine the dynamic analysis of user-role and topic-influence with probability model to predict topic propagation trends. Third, we present the pseudocode of the TPP-DA algorithm.

A. THE PROBABILITY MODEL

User behaviors, relationships and time spans are used to calculate the probability of user engagement on a topic. Specifically, the number of times a user participates in a topic, the number of people participating in a topic and the popularity of the topic will affect the user's behavior in different ways. In other words, a probability model is built according to behavior probability, relationship probability, and time probability.

1) BEHAVIOR PROBABILITY

If a user often participates in the discussion of a certain topic, then he or she is interested in the topic and may discuss the topic in the future. As a result, we designate a behavior probability function $p_b(u_i, z, t)$ for user u_i engaging in topic z at time t, which is:

$$p_b(u_i, z, t) = p_b(x_i^z(t)) = \prod_{n=t-q}^{t-1} (k_z^{[2-(x_i^z(t)-x_i^z(n))^2]} \cdot e^{-k_z}) \tag{13}$$

where $x_i^z(n)$ denotes whether user u_i discusses topic z at time n, k_z is a parameter that is trained for topic z in advance, e is the base of the natural logarithms, and q means that the user's behavior probability at time t is associated with the behavior probabilities from time $(t - q)$ to time $(t - 1)$. When $x_i^z(n)$ is equal to 1, the value of the behavior probability function $p_b(u_i, z, t)$ is directly proportional to the probability of user u_i joining the discussion at time n. Conversely, it is directly proportional to the probability of user u_i not joining the discussion when $x_i^z(n) = 0$.

2) RELATIONSHIP PROBABILITY

Considering a user's social relationships, if more users or friends participate in a discussion, then he or she will be more willing to discuss the topic. We designate a relationship probability function $p_r(u_i, z, t)$ for user u_i who discusses topic z at time t. The function $p_r(u_i, z, t)$ is:

$$p_r(u_i, z, t) = p_r(x_i^z(t)) = \prod_{n=t-q'}^{t-1} (l_z^{[1+y_z^{(x_i^z(t)+x_i^z(n)-1)^2}]} \cdot e^{-l_z}) \tag{14}$$

$$y_z = \frac{(\sum_{i=1}^{N} \sum_{n=t-(q'-1)}^{t-1} x_i^z(n)) + 1}{(\sum_{i=1}^{N} \sum_{n=t-q'}^{t-2} x_i^z(n)) + 1} \tag{15}$$

where y_z is the ratio of the change of the user number, which is calculated by dividing the total number of users involved in the discussion of topic z from time $(t - (q' - 1))$ to $(t - 1)$ by the total number of those involved in the discussion of topic z from time $(n = t - q')$ to $(t - 2)$. If the total number of users decreases during the time period, y_z is less than 1. In addition, q' means that a user's relationship probability at time t is related to the relationship probabilities from time $(t - q')$ to time $(t - 1)$. In addition, l_z is a parameter that is trained for topic z in advance, and e is the base of the natural logarithms.

3) TIME PROBABILITY

Regarding the effectiveness of user interests, we assume that users will gradually lose interest after participating in topic discussions for a period. Generally, the total number of participants decreases after reaching the peak number of participants, and the rate of decline increases with the period of time after the peak time. We designate a time probability function $p_t(u_i, z, t)$ for user u_i engaging in topic z at time t to calculate the probability using a time lapse factor:

$$p_t(u_i, z, t) = p_t(x_i^z(t)) = \frac{1}{(t - t_m^z)^{\lambda_z \cdot x_i^z(t)}} \tag{16}$$

where t_m^z is the peak time when the number of participants is the highest from the initial time to time t on topic z, and λ_z is a lapse exponential coefficient to be evaluated by experience. If $x_i^z(t) = 1$, then the value of the time probability function is larger than 1; and the greater t is, the lower the value of the function will be. Thus, the longer the interval from the peak time to the predicted time is, the lower the probability of a user joining the discussion will be. If $x_i^z(t) = 0$, then the value of the time probability function is 1.

B. TOPIC PROPAGATION MODEL

The dynamic user-role and topic-influence analysis are combined with the probability model and integrated into a unified topic propagation prediction model.

First, according to the analysis, users have different roles in different topics, and the user-role influences also change over time. Dynamic user-role analysis should be considered in topic propagation. Second, the topic itself plays an important role in topic propagation, therefore topic-influence should be considered in topic propagation. The user-role (the user

leader-factor and expert-factor) and topic-influence are combined in a single user with the behavior probability to obtain the following weighted probability:

$$PB(u_i, z, t) = W_1 \cdot p_b(u_i, z, t) \tag{17}$$

$$W_1 = [F_{ep}(u_i, z, t) + F_{ld}(u_i, z, t)] \cdot G_{tu}(z, u_i, t) \tag{18}$$

Similarly, the user social relationships (i.e., the social-factor and similarity-factor) and topic-influence on social groups are combined with the relationship probability to obtain:

$$PR(u_i, z, t) = W_2 \cdot p_r(u_i, z, t) \tag{19}$$

$$W_2 = [\sum(F_{sc}(u_i, u_x, z, t) + F_{sl}(u_i, u_x, z, t))] \\ \cdot G_{tg}(z, U, t) \tag{20}$$

As time passes, the probability of a particular user engaging in a specific topic discussion decreases. Moreover, this probability is affected by the user's roles. In general, if a user is an expert user on a topic, he or she will continue to pay attention to it. Otherwise, he or she may lose interest in the topic after a period of time. Thus, we revise the time probability of a user on a topic as:

$$PT(u_i, z, t) = W_3 \cdot p_t(u_i, z, t) \tag{21}$$

$$W_3 = \frac{F_{ex}(u_i, z, t)}{\max_x(F_{ex}(u_x, z, t))} \tag{22}$$

Considering user-role and topic-influence analysis and the weighted probability model together, we can devise a behavioral tendency function, $\Omega(u_i, z, t)$, for user u_i on topic z at time t, which is denoted as:

$$\begin{aligned}
P(x_i^z(t)) = P(u_i, z, t) &\propto \Omega(x_i^z(t)) = \Omega(u_i, z, t) \\
&= PB(u_i, z, t) \cdot PR(u_i, z, t) \cdot PT(u_i, z, t) \\
&= W \cdot p_b(u_i, z, t) p_r(u_i, z, t) p_t(u_i, z, t) \\
&= W \cdot p_b(x_i^z(t)) p_r(x_i^z(t)) p_t(x_i^z(t)) \tag{23}
\end{aligned}$$

Here, $P(x_i^z(t))$ is the probability of the tendency of user u_i to engage in the discussion of topic z at time t. $P(x_i^z(t))$ has a positive correlation with $\Omega(x_i^z(t))$. $W = W_1 \cdot W_2 \cdot W_3$.

Based on the above equations, the parameters k_z, l_z, and λ_z can be estimated for each topic z by the sample date. Then, we can calculate $\Omega(x_i^z(t))$ by setting $x_i^z(t)$ to 1 or 0. If $x_i^z(t) = 1$, $\Omega(x_i^z(t))$ has a positive correlation with $P(x_i^z(t) = 1)$; and if $x_i^z(t) = 0$, $\Omega(x_i^z(t))$ has a positive correlation with $P(x_i^z(t) = 0)$. After normalizing $\Omega(x_i^z(t))$, we can obtain the values of $P(x_i^z(t) = 1)$ and $P(x_i^z(t) = 0)$. If $P(x_i^z(t) = 1) \geq P(x_i^z(t) = 0)$, we assume that user u_i is willing to engage in topic z at time t, i.e., $x_i^z(t) = 1$; otherwise, we assume that user u_i will not engage in topic z at time t, i.e., $x_i^z(t) = 0$.

C. ALGORITHMS

Because the TPP-DA is based on the dynamic analysis of user-role and topic-influence, we first should train the model

and obtain the user-role factors and topic-influences. Based on the training and testing results, we provide the details of the TPP-DA algorithm.

We sort all messages from time 1 to time T in chronological order, and then we divide them into a training set and a testing set, which is $M_{train} = \{M_{train}(1), M_{train}(2), \ldots, M_{train}(t)\}$ and $M_{test} = \{M_{test}(t + 1), M_{test}(t + 2), \ldots, M_{test}(T)\}$. Generally, the training set and testing set are independent, i.e., $M_{train} \cap M_{test} = \emptyset$ and $M_{train} \cup M_{test} = M_{all}$. To correctly train user-roles and improve the prediction accuracy, the training set and testing set must cover all users and topics, i.e., $U_{test} = U_{all}$ and $C_{train} = C_{test} = C_{all}$.

1) DYNAMIC USER-ROLE TRAINING AND TESTING

The pseudocode of the user-role training algorithm is depicted in algorithm 1. First, the training set is sorted in chronological order and divided into multiple subsets in lines 1-2. For each subset, we update some information (lines 5-9) and then calculate the values of the user-roles (lines 10-17). Finally, we return the training results from time 1 to time t in line 19.

Algorithm 1 Dynamic User-Role Training Algorithm

Input: Training set M_{train}
Output: User-role training result $F_{n=1}^t$
1: Sort M_{train} in chronological order;
2: Divide M_{train} into subsets $M_{train}(1), M_{train}(2), \ldots, M_{train}(t)$;
3: **for** $(n = 1, n <= t, n++)$ **do**
4:     Get subset $M_{train}(n)$;
5:     **for** each message $m \in M_{train}(n)$ **do**
6:       **for** each topic $z \in C$ **do**
7:         Update $M_i^z(n)$ of user u_i;
8:       **end for**
9:     **end for**
10:     **for** each topic $z \in C$ **do**
11:       **for** user $u_i \in U$ **do**
12:         Calculate $F_{ep}(u_i, z, n)$ and $F_{ld}(u_i, z, n)$;
13:         **for** each user $u_{j(j \neq i)} \in U$ **do**
14:           Calculate $F_{sc}(u_i, u_j, z, n)$ and $F_{sl}(u_i, u_j, z, n)$;
15:         **end for**
16:       **end for**
17:     **end for**
18: **end for**
19: **return** $F_{n=1}^t = \{F_{ep}, F_{ld}, F_{sc}, F_{sl}\}_{n=1}^t$

Algorithm 2 predicts user-roles at unknown times. The testing set is also sorted and divided into multiple subsets (lines 1-2). For each subset, we construct the Lagrange formulas and predict the values of user-roles at time $(t + k)$ (lines 5-13). Then, we update $x_i^z(t + k)$ at time $(t + k)$ with the subset, which is prepared for the next iteration at time $(t + k + 1)$. Finally, we return the testing results from time $(t + 1)$ to time T in line 16.

Algorithm 2 Dynamic User-Role Testing Algorithm

Input: User-role training result $F_{n=1}^t$, testing set M_{test}, s
Output: User-role analysis results $F_{n=t+1}^T$
1: Sort M_{test} in chronological order;
2: Divide M_{test} into multiple subsets $M_{test}(t+1)$, $M_{test}(t+2)$, ..., $M_{test}(T)$;
3: **for** $(k = 1, k <= T, k++)$ **do**
4:     Get subset $M_{test}(t+k)$;
5:     **for** each topic $z \in C$ **do**
6:     **for** user $u_i \in U$ **do**
7:     Get the previous s training results;
8:     Construct Lagrange formulas L_{ep} and L_{ld} and calculate $F_{ep}(u_i, z, n+k)$ and $F_{ld}(u_i, z, t+k)$;
9:     **for** each user $u_{j(j \neq i)} \in U$ **do**
10:     Construct Lagrange formulas L_{sc} and L_{sl}, and calculate $F_{sc}(u_i, u_j, z, t+k)$ and $F_{sl}(u_i, u_j, z, t+k)$;
11:     **end for**
12:     **end for**
13:     **end for**
14:     Update $x_i^z(t+k)$ at time $(t+k)$ with $M_{test}(t+k)$;
15: **end for**
16: **return** $F_{n=t+1}^T = \{F_{ep}, F_{ld}, F_{sc}, F_{sl}\}_{n=t+1}^T$

Algorithm 3 Dynamic Topic-Influence Training Algorithm

Input: Training set M_{train}
Output: Topic-influence training result $G_{n=1}^t$
1: Sort M_{train} in chronological order;
2: Divide M_{train} into multiple subsets $M_{train}(1)$, $M_{train}(2)$, ..., $M_{train}(t)$;
3: **for**$(n = 1, n <= t, n++)$ **do**
4:     Get subset $M_{train}(n)$;
5:     **for** each message $m \in M_{train}(n)$ **do**
6:     **for** each topic $z \in C$ **do**
7:     Update $M_i^z(n)$ of user u_i;
8:     **end for**
9:     **end for**
10:     **for** each topic $z \in C$ **do**
11:     **for** user $u_i \in U$ **do**
12:     Calculate $G_{tu}(z, u_i, n)$;
13:     **end for**
14:     Calculate $G_{tg}(z, u_i, n)$;
15:     **end for**
16: **end for**
17: **return** $G_{n=1}^t = \{G_{tu}, G_{tg}\}_{n=1}^t$;

Next, we analyze the computational complexity of algorithms 1 and 2. If there are N users, K topics, and M messages, Algorithm 1 first iterates M times in the "for" loop (lines 3-18). The computational complexities of lines 5-9 and lines 10-17 are $O(K \times M)$ and $O(K \times M \times N)$ respectively. As a result, the computational complexity of algorithm 1 is $O(K \times M \times N^2)$. Generally, there are few topics, thus K is a small fixed value, and the computational complexity of algorithm 1 can be $O(M \times N^2)$. Similarly, algorithm 2 has the same computational complexity as $O(M \times N^2)$.

2) DYNAMIC TOPIC-INFLUENCE TRAINING AND TESTING
The processes of topic-influence training and testing are similar to those of user-role training and testing. Algorithms 3 and 4 give the training and testing processes of the topic-influence analysis respectively. In Algorithm 3, we first update some information for each subset (lines 5-9), and then calculate the topic-influences (lines 10-15). In Algorithm 4, we construct the Lagrange formulas and predict the values of topic-influences at time $(t + k)$ (lines 5-11). Next, we update $x_i^z(t+k)$ with the subset, which is prepared for the next iteration at time $(t + k + 1)$.

In Algorithm 3, the computational complexities of lines 5-9 and lines 10-15 are $O(K \times M)$ and $O(K \times N)$ respectively. Moreover, the complexity of the loop (line 3) is $O(M)$. Therefore, the computational complexity of Algorithm 3 is $O(K \times M \times N)$. Since K is a fixed value, the complexity is $O(M \times N)$. The complexity of Algorithm 4 is also $O(M \times N)$.

Algorithm 4 Dynamic Topic-Influence Testing Algorithm

Input: Topic-influence training result $G_{n=1}^t$, testing set M_{test}, and s
Output: Topic-influence analysis results $G_{n=t+1}^T$
1: Sort M_{test} in chronological order;
2: Divide M_{test} into multiple subsets $M_{test}(t+1)$, $M_{test}(t+2)$, ..., $M_{test}(T)$;
3: **for** $(k = 1, k <= T, k++)$ **do**
4:     Get subset $M_{test}(t+k)$;
5:     **for** each topic $z \in C$ **do**
6:     **for** user $u_i \in U$ **do**
7:     Get the previous s results;
8:     Construct Lagrange formula L_{tu} and calculate $G_{tu}(z, u_i, t+k)$;
9:     **end for**
10:     Construct Lagrange formula L_{tg} and calculate $G_{tg}(z, U, t+k)$;
11:     **end for**
12:     Update $x_i^z(t+k)$ with $M_{test}(t+k)$;
13: **end for**
14: **return** $G_{n=t+1}^T = \{G_{tu}, G_{tg}\}_{n=t+1}^T$

---

3) TPP-DA ALGORITHM
The TPP-DA algorithm is depicted in Algorithm 5. According to the previous subsections, $p_b(u_i, z, n)$ and $p_r(u_i, z, n)$ are associated with the previous q and q' values of user u_i, and $p_t(u_i, z, n)$ is associated with the peak time on topic z. At time n, the previous q and q' values can be obtained. Then, we can calculate the current $p_b(u_i, z, n)$ and $p_r(u_i, z, n)$ with Eqs. (13) and (14) respectively. By updating

---

the peak time t_m, $p_t(u_i, z, n)$ can be calculated with Eq. (16). Then, the prediction value $\Omega(x_i^z(n))$ is calculated with Eq. (23). After normalizing $\Omega(x_i^z(n))$, we can predict the value of $x_i^z(n)$ and calculate the number of users discussing topic z at time n. Last, the test date is used to update the real $x_i^z(n)$ and t_m and to prepare for the next iteration.

Algorithm 5 TPP-DA Algorithm

Input: User-role factors $F_{n=1}^T$, topic-influences $G_{n=1}^T$, M_{test}, q, and q'
Output: Number of users on topic z at time n: $N(n, z)$
1: Sort messages in chronological order;
2: Divide M_{test} into multiple subsets by time, $M_{test}(t + 1)$, $M_{test}(t + 2)$, ..., $M_{test}(T)$;
3: $n = t + 1$;
4: **while** $n + + \leq T$ **do**
5:   **for** each topic $z \in C$ **do**
6:     **for** each user $u_i \in U$ **do**
7:       Get the peak time t_m on topic z;
8:       Calculate $p_b(u_i, z, n)p_r(u_i, z, n)p_t(u_i, z, n)$;
9:       Calculate $\Omega_1(x_i^z(n))$ and $\Omega_0(x_i^z(n))$ when $x_i^z(n) = 1$ and $x_i^z(n) = 0$, respectively;
10:    **end for**
11:    Normalize $\Omega_1(x_i^z(n))$ and $\Omega_0(x_i^z(n))$ for all users;
12:    **for** each user $u_i \in U$ **do**
13:      $x_i^z(n) = 0$;
14:      Get $P(x_i^z(n) = 1)$ and $P(x_i^z(n) = 0)$;
15:      **if** $P(x_i^z(n) = 1) \geq P(x_i^z(n) = 0)$ **then**
16:        $x_i^z(n) = 1$;
17:      **end if**
18:    **end for**
19:    Calculate $N(n, z) = \sum_{i=1}^N x_i^z(n)$;
20: **end for**
21: Update $x_i^z(n)$ and t_m^z with $M_{test}(n)$;
22: **end while**
23: **return** $N(n, z)$;

In Algorithm 5, the computational complexities of lines 6-10 and lines 12-18 are both $O(N)$. The statements in the loop bodies (lines 5-20 and lines 4-22) run K and M times, respectively. Thus, the computational complexity of Algorithm 5 is $O(K \times M \times N)$. Finally, since K is a fixed value, the computational complexity is $O(M \times N)$.

VI. EXPERIMENTS

A. DATASET

To evaluate the effectiveness of the TPP-DA method, suitable social network datasets are needed. The datasets are collected from Sina Weibo and Twitter. After crawling the data, the Sina dataset and Twitter dataset are obtained. The Sina dataset contains 8586 users, 416826 microblogs (messages), and 98362 user relationships, while the Twitter dataset has 6060 users, 15485 tweets (messages), and 11617 user relationships. These data are managed in three tables: a user table, a blog table, and a user relationship table. The user

table contains user information; the blog table contains all attributes of all messages; and the relationship table consists of two fields, "suid" and "tuid", indicating that "suid" pays attention to "tuid".

As there are many topics in social networks, we focus on only hot topics for the sake of simplicity. In the Sina dataset, four hot topics are chosen: "US Presidential Elections", "Novel Coronavirus Pneumonia", "Smog", and "Elon Musk", which are labeled as the 1-1st topic, 1-2nd topic, 1-3rd topic and 1-4th topic, respectively. In the Twitter dataset, two topics, which are "tokyo_2020" and "capitol_2020" are selected and labeled as 2-1st topic and 2-2nd topic, respectively. After data preprocessing, we remove messages that do not relate to the above topics. Then, the datasets are divided into six sub-datasets. For example, 19458 microblogs and 5472 relationships among 1228 users are related to the 1-1st topic "US Presidential Elections" in the first sub-dataset, 17184 microblogs and 5164 relationships among 1305 users are related to the 1-2nd topic "Novel Coronavirus Pneumonia" in the second sub-dataset, and so on. Last, we obtain the Sina dataset with 50127 microblogs and the Twitter dataset with 12367 tweets. The Sina dataset is related to 3672 users and four topics, and there are 14846 relationships among these 3672 users, while the Twitter dataset contains 982 users, 7636 relationships and 12367 tweets related to two topics. The detailed information of the datasets is shown in Table 2 and Table 3.

TABLE 2. The detailed information of Sina dataset.

	Topic	users	relationships	blogs
1-1st topic	US Presidential Elections	1228	5472	19458
1-2nd topic	Novel coronavirus pneumonia	1305	5164	17184
1-3rd topic	Smog	862	2734	9081
1-4th topic	Elon Musk	1216	4083	13561
1-all topics	All topics	3672	14846	50127

TABLE 3. The detailed information of Twitter dataset.

	Topic	users	relationships	tweets
2-1st topic	tokyo_2020	483	3928	5643
2-2nd topic	capitol_2020	664	5362	7835
2-all topics	All topics	982	7636	12367

Note that each user may have discussed more than one topic, and the sum of the users on all topics may be larger than the number of users in the dataset, which is similar for the numbers of relationships and messages. All messages are sorted in chronological order into two datasets. Then, the datasets are divided into a training set and testing set. The training set and testing set are independent, and both of them cover all users and topics in each dataset. To show

the dynamic analysis of user-role and topic-influence, the training set and testing set were divided into subsets for every day. The training set covers messages from time 1 to t, and the testing set contains the latter messages ranging from $(t + 1)$ to T. Here, the time unit in $[1, T]$ is one day.

B. PARAMETER SETTINGS

1) PARAMETER OF THE LAGRANGE FORMULA TRAINING

The first important parameter used to construct the Lagrange interpolation formula in the dynamic analysis of user-role and topic-influence is s. As mentioned in Eq. (11), we can predict the next value when we know the previous s values; the value of parameter s affects the Lagrange interpolation formula. Generally, the larger the parameter s is, the more complex the constructed Lagrange formulas will be. As a result, we cannot set parameter s too large. In our experiments, we set s as 2, 3, 5, 7, and 9. Then, we construct the Lagrange formula in Eq. (11) to dynamically predict user expert-factors from time $(t + 1)$ to T with Eq. (12).

After constructing the Lagrange formula with the previous s values, we calculate the predicted expert-factors with Eq. (12). Next, we compare them to the results calculated with Eq. (1) using the actual data. The average error rates between them with s varying are shown in Table 4.

TABLE 4. Average error rates of the predicted expert-factor values of the Lagrange formula and the actual expert-factor values of the dynamic user-role analysis with *s* varying on different topics.

s	2	3	5	7	9
1-1st topic	21.3	21.6	13.7	13.2	15.3
1-2nd topic	25.6	22.8	16.5	17.1	17.6
1-3rd topic	25.2	25.1	15.3	16.3	17.9
1-4th topic	24.4	24.7	19.7	19.4	19.8
2-1st topic	23.6	22.5	16.2	16.9	18.3
2-2nd topic	23.2	22.1	16.0	16.4	18.0

Table 4 shows that the average error rates are the smallest when s is set as 5 or 7 on different topics. When $s = 2$ or $s = 3$, the Lagrange formulas will be linear functions and quadratic polynomial functions. In these settings, the average error rates are very large because they are not suitable for sudden changes in user expert-factor values. The constructed Lagrange formulas are too complex when s is set as 9, and the average error rates are also large, which may be due to overfitting. The results of setting $s = 5$ and $s = 7$ are the same. We set $s = 5$ to construct the Lagrange formulas when analyzing the user expert-factor dynamically.

As mentioned previously, the user expert-factor and similarity-factor are content-based factors, and the leader-factor and social-factor are relationship-based factors; therefore, we take the user leader-factor as an example to conduct a dynamic user-role analysis experiment when s is set as 2, 3, 5, 7, and 9. The results are listed in Table 5. As analyzed in Section IV, the user leader-factor is calculated based on the

TABLE 5. Average error rates of the predicted leader-factor values of the Lagrange formula and the actual leaders-factor values of the dynamic user-role analysis with s varying on different topics.

s	2	3	5	7	9
1-1st topic	9.42	8.4	9.22	8.86	10.22
1-2nd topic	9.13	8.76	9.30	10.73	8.31
1-3rd topic	9.85	8.31	9.11	8.64	10.52
1-4th topic	9.95	8.9	9.64	9.83	9.96
2-1st topic	9.26	9.03	9.16	9.67	9.82
2-2nd topic	9.18	8.96	9.24	9.53	9.65

relationships, which are generally stable in social networks; therefore, the change in the user leader-factor is very small when s varies. For simplicity, we set $s = 5$ when constructing the Lagrange formulas to predict the leader-factor value.

Last, we test the average error rates of topic-influence analysis, and the results are similar. In summary, we set $s = 5$ in the dynamic analysis of user-role and topic-influence.

2) PARAMETERS OF THE PROBABILITY MODEL

In addition to the parameter s mentioned above, q, q' and λ are used to calculate the values in Eqs. (13), (14), and (16), respectively. In our previous work, the 1-2nd topic and 1-4th topic were used as examples to calculate the average errors between the prediction results and actual data with varying settings of parameters q, q' and λ, and the results are shown in Table 6. According to the experimental results, the model works best when $q = q' = 3$ and $\lambda = 0.5$. We also set $q = q' = 3$ and $\lambda = 0.5$ as the default values in our experiments.

TABLE 6. Average errors of the probability model on different topics.

λ		0.3	
q, q'	1	3	5
1-2nd topic	16.7	14.5	17.8
1-4th topic	24.4	19.9	26.2
λ		0.5	
q, q'	1	3	5
1-2nd topic	15.4	12.2	14.8
1-4th topic	21.3	17.6	20.3
λ		1	
q, q'	1	3	5
1-2nd topic	19.6	16.5	21.4
1-4th topic	25.6	22.7	28.9

C. EXPERIMENTAL RESULTS

1) COMPARISON RESULTS WITH OUR PREVIOUS WORK

In this subsection, the TPP-DA method is evaluated by comparing the prediction results of TPP-DA against those of our previous work (TPP). Furthermore, we compare the results of TPP-DA with actual data. We take time (each day) as the abscissa axis and the number of users who may discuss the topic at time n as the ordinate axis to draw curves. The prediction results of TPP-DA and TPP against the actual data on all topics are shown in Fig. 4. Moreover, we calculate
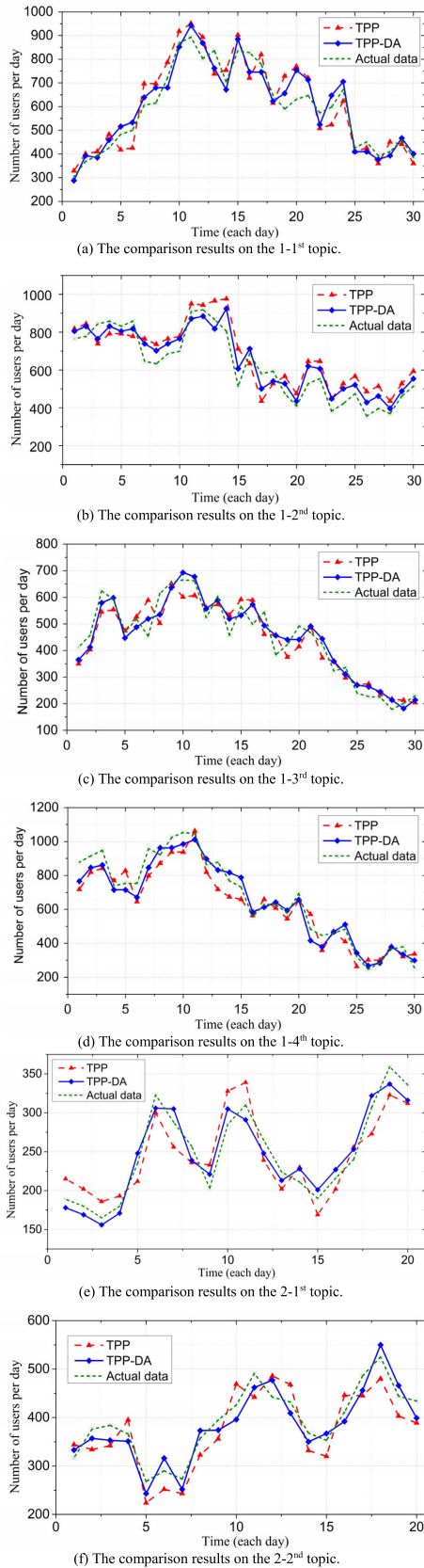
(a) The comparison results on the 1-1st topic.

(b) The comparison results on the 1-2nd topic.

(c) The comparison results on the 1-3rd topic.

(d) The comparison results on the 1-4th topic.

(e) The comparison results on the 2-1st topic.

(f) The comparison results on the 2-2nd topic.

FIGURE 4. The results of the TPP, TPP-DA and actual data on each day.

the average error rates between the prediction results and the actual data, which are shown in Table 7.

Fig. 4 and Table 7 show that both TPP-DA and TPP can predict topic propagation trends accurately. Moreover, TPP-DA outperforms TPP on all topics in two datasets. In particular, Table 7 shows that all average error rates of TPP are over 10%, and even the error rate of TPP on the 1-2nd topic is nearly 16%. All average error rates of TPP-DA are less than 10%, which indicates that TPP-DA has a smaller error rate. Compared with TPP, the average error rate of TPP-DA is reduced by approximately 33% for two reasons. The first is that TPP-DA utilizes dynamic user-role analysis, which is more adaptable to user-roles changing over time in social networks. Conversely, TPP analyzes user-role only statically, which ignores that user-roles are time dependent. The second is that TPP-DA introduces topic-influence analysis. TPP-DA considers the competition among topics to analyze topic-influences on both single user and social group, so it can model topic propagation trends more accurately. To summarize, TPP-DA is based on the dynamic analysis of user-role and topic-influence. Thus, it has better results than TPP, which also proves the advantage of dynamic analysis.

TABLE 7. Average error rates between the prediction results and actual data.

	1-1st topic	1-2nd topic	1-3rd topic	1-4th topic	2-1st topic	2-2nd topic
TPP	10.14	15.98	11.55	10.99	10.21	10.05
TPP-DA	7.18	9.85	8.84	6.85	6.45	6.97

In addition, both TPP-DA and TPP incorporate behavior probability, relationship probability and time probability to build a weighted probability model that predicts topic propagation. They can handle multiple peaks in the topic propagation process. For example, there are approximately four peaks at days 11, 15, 21, and 24 on the spreading of the 1-1st topic. At these peaks, both TPP-DA and TPP can obtain similar prediction topic propagation trends. Because TPP-DA obtains the weights of the probability model based on the dynamic analysis of user-role and topic-influence, it can adjust these weights with time to adapt to the changeability dynamically. Therefore, TPP-DA has more accurate results than TPP. The experimental results in Fig. 4 and Table 7 show the effectiveness of TPP-DA on topic propagation prediction in social networks.

2) COMPARISON RESULTS WITH OTHER METHOD

In this subsection, TPP-DA is compared with another information diffusion method called temporal dynamics of information diffusion (TDID) [20], which was also compared with TPP in our previous work. According to [20], the total number
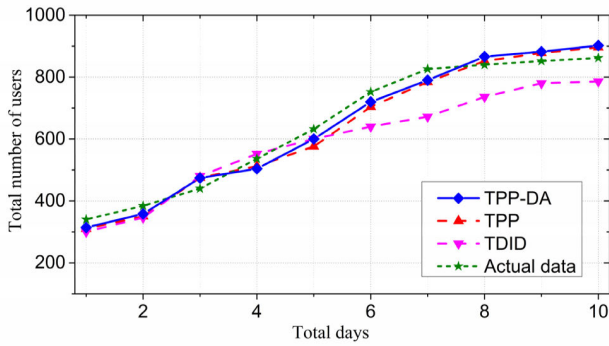
FIGURE 5. The comparison results of TPP, TPP-DA and TDID on 1-3rd topic.

of infected users at time t can be obtained by:

$$I(t) = N + \frac{B(N - I(1))t^B}{-B - A(N - I(t)) + A(N - I(t))t^B} \quad (24)$$

where $A = (K_{avg}^{out}\lambda_1(t) - \lambda_2(t))/N(t)$ and $B = -(K_{avg}^{out}\lambda_1(t) + \lambda_2(t))$. K_{avg}^{out} is the average out-degree of users, and $\lambda_1(t)$ and $\lambda_2(t)$ denote the probability/rate that an infected or susceptible user would publish a tweet with the particular hashtag of interest.

In this experiment, we use the same parameter settings used in our previous work. That is, $N = 862$ (the number of users on the 1-3rd topic), $\lambda_1(t) = \frac{0.2}{1 + 1/t}$, $\lambda_2(t) = \frac{0.1}{1 + 1/t}$, and $K_{avg}^{out} = 4$.

Furthermore, we record the set of users who discuss the topic at each time t as $N_t \in U$. Next, we aggregate them as $N_{total} = \bigsqcup_1^t N_t$ to obtain the total number of users discussing the topic from the start time to time t. The actual results are obtained by analyzing the date set. The comparison results of TPP-DA, TPP, TDID and the actual data on 1-3rd topic are shown in Fig. 5.

The actual total number of infected users reaches its maximum (862 users) on the 10th day, and it will not change over time; therefore, the results for only the first 10 days are shown. TPP-DA performs the best among the three methods, and TPP has the second-best results. This is because both TPP-DA and TPP consider user-role analysis in topic propagation, while TDID develops only an epidemic model of the temporal dynamics of information propagation for specific topics, which proves that user-role analysis is helpful for topic propagation. Moreover, TPP-DA performs slightly better than TPP, which shows that the dynamic analysis of user-role and topic-influence used in TPP-DA is better than the static user-role analysis used in TPP.

In addition, we compare TPP-DA and TPP with TDID using the MAE (mean absolute error), RMSE (root mean square error) and R^2 (R-squared, i.e., the coefficient of determination), and the results are listed in Table 8. TPP-DA has the smallest average errors (MAE and RMSE) and the largest R^2. For example, the R-squared of TPP-DA is 0.948 on the 1-1st topic, which is higher than those of TPP and TDID,

TABLE 8. Average error rates of prediction results of TPP-DA, TPP and TDID of different topics.

Methods	1-1st topic			1-2nd topic		
	MAE	RMSE	R^2	MAE	RMSE	R^2
TPP-DA	4.8	5.65	0.948	5.1	5.01	0.937
TPP	5.6	6.45	0.930	5.2	5.18	0.930
TDID	11.9	12.68	0.869	10.8	15.2	0.846
Methods	1-3rd topic			1-4th topic		
	MAE	RMSE	R^2	MAE	RMSE	R^2
TPP-DA	5.3	5.53	0.974	4.2	4.05	0.951
TPP	5.9	6.42	0.966	4.6	5.22	0.940
TDID	10.2	11.07	0.835	11.6	12.01	0.823
Methods	2-1st topic			2-2nd topic		
	MAE	RMSE	R^2	MAE	RMSE	R^2
TPP-DA	4.16	4.44	0.969	3.52	4.23	0.938
TPP	9.04	9.41	0.913	4.89	5.33	0.886
TDID	11.53	11.57	0.824	5.91	6.34	0.805

and the MAE and RMSE of TPP-DA are smaller than those of TPP and TDID. These results mean that TPP-DA is closest to the actual results on topic propagation, which also proves that introducing dynamic analysis of user-role and topic-influence is more suitable for topic propagation prediction in social networks.

VII. CONCLUSION

In this paper, we analyzed the shortcomings of both the existing works and our previous work. Then, we applied dynamic analysis of user-role and topic-influence to propose the TPP-DA method, which can predict topic propagation considering both users' and topics' perspectives. First, TPP-DA introduces a temporal perspective to static user-role analysis to analyze user-role dynamically and accurately, which is helpful for topic propagation. Second, the topic competition and topic-influences in social networks are considered when studying topic propagation. TPP-DA also utilizes a metric called the topic heat to calculate the topic-influences on a single user and social group. Third, TPP-DA combines the dynamic analysis of user-role and topic-influence with a weighted probability model to predict topic propagation. Last, two datasets are crawled from Sina and Twitter, and experiments are conducted to evaluate TPP-DA on these two datasets. The experimental results show that TPP-DA performs more effectively and efficiently than others. Moreover, the average error rate of TPP-DA is 33% lower than that of the TPP, which also proves the effectiveness of the dynamic analysis of user-role and topic-influence in topic propagation.

Although the TPP-DA can get better results than the others, it is a probability-based prediction model. In the future, we will consider how to utilize deep learning method to predict topic propagation. Moreover, we also should crawl more data to verify our method.

REFERENCES

[1] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, "Rumor propagation dynamic model based on evolutionary game and anti-rumor," *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, 2019.

[2] Y. Xiao, Q. Yang, C. Sang, and Y. Liu, "Rumor diffusion model based on representation learning and anti-rumor," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 3, pp. 1910–1923, Sep. 2020.

[3] Y. P. Xiao, W. Li, S. Qiang, Q. Li, and Y. Liu, "A rumor & anti-rumor propagation model based on data enhancement and evolutionary game," *IEEE Trans. Emerg. Topics Comput.*, early access, Oct. 27, 2020, doi: 10.1109/TETC.2020.3034188.

[4] Z. Wang, C. Chen, and W. Li, "Attention network for information diffusion prediction," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 65–66.

[5] Z. Bu, H.-J. Li, C. Zhang, J. Cao, A. Li, and Y. Shi, "Graph K-means based on leader identification, dynamic game, and opinion dynamics," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 7, pp. 1348–1361, Jul. 2020.

[6] Y. Y. Wang, J. Cao, Z. Bu, J. Jiang, and H. Chen, "Proximity-based group formation game model for community detection in social network," *Knowl.-Based Syst.*, vol. 214, pp. 1–13, Feb. 2021.

[7] Z. Bu, Y. Y. Wang, H.-J. Li, J. Jiang, Z. Wu, and J. Cao, "Link prediction in temporal networks: Integrating survival analysis and game theory," *Inform. Sci.*, vol. 498, pp. 41–61, Sep. 2019.

[8] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn, "Optimal network modularity for information diffusion," *Phys. Rev. Lett.*, vol. 113, no. 8, pp. 1–5, Aug. 2014.

[9] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proc. 21st Int. Conf. Companion World Wide Web (WWW) Companion*, 2012, pp. 1145–1152.

[10] C. Gray, L. Mitchell, and M. Roughan, "Super-blockers and the effect of network structure on information cascades," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 1435–1441.

[11] Z. Wang, C. Chen, and W. Li, "Information diffusion prediction with network regularized role-based user representation learning," *ACM Trans. Knowl. Discovery Data*, vol. 13, no. 3, pp. 1–23, Jul. 2019.

[12] Y. Zhou, B. Zhang, X. Sun, Q. Zheng, and T. Liu, "Analyzing and modeling dynamics of information diffusion in microblogging social network," *J. Netw. Comput. Appl.*, vol. 86, pp. 92–102, May 2017.

[13] S. Gao, H. Pang, P. Gallinari, J. Guo, and N. Kato, "A novel embedding method for information diffusion prediction in social network big data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2097–2105, Aug. 2017.

[14] X. Rui, F. Meng, and Z. Wang, "SPIR: The potential spreaders involved SIR model for information diffusion in social networks," *Phys. A, Stat. Mech. Appl.*, vol. 506, pp. 254–269, Sep. 2018.

[15] Y. Yi, Z. Zhang, L. T. Yang, C. Gan, X. Deng, and L. Yi, "Reemergence modeling of intelligent information diffusion in heterogeneous social networks: The dynamics perspective," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 828–840, Apr./Jun. 2021.

[16] Q. Kong, M.-A. Rizoiu, and L. Xie, "Modeling information cascades with self-exciting processes via generalized epidemic models," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 286–294.

[17] K. Nagata and S. Shirayama, "Analysis method of influence of potential edge on information diffusion," *Proc. Comput. Sci.*, vol. 4, pp. 241–250, Jan. 2011.

[18] S. Molaei, H. Zare, and H. Veisi, "Deep learning approach on information diffusion in heterogeneous networks," *Knowl.-Based Syst.*, vol. 189, pp. 1–13, Feb. 2020.

[19] Q. Li, Z. Wang, B. Wu, and Y. Xiao, "Competition and cooperation: Dynamical interplay diffusion between social topic multiple messages in multiplex networks," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 3, pp. 467–478, Jun. 2019.

[20] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, "Temporal dynamics of information diffusion in twitter: Modeling and experimentation," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 256–264, Mar. 2018.

[21] X. Liu, D. He, and C. Liu, "Information diffusion nonlinear dynamics modeling and evolution analysis in online social network based on emergency events," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 1, pp. 8–19, Feb. 2019.

[22] X. Cao, Y. Chen, C. Jiang, and K. J. Ray Liu, "Evolutionary information diffusion over heterogeneous social networks," *IEEE Trans. Signal Inf. Process. over Netw.*, vol. 2, no. 4, pp. 595–610, Dec. 2016.

[23] M. Farajtabar, M. Gomez-Rodriguez, Y. Wang, S. Li, H. Zha, and L. Song, "COEVOLVE: A joint point process model for information diffusion and network co-evolution," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 473–477.

[24] Y. Hu, R. J. Song, and M. Chen, "Modeling for information diffusion in online social networks via hydrodynamics," *IEEE Access*, vol. 5, pp. 128–135, 2017.

[25] A. Saha, B. Samanta, N. Ganguly, and A. De, "CRPP: Competing recurrent point process for modeling visibility dynamics in information diffusion," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 537–546.

[26] X. Chen, K. Zhang, F. Zhou, G. Trajcevski, T. Zhong, and F. Zhang, "Information cascades modeling via deep multi-task learning," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 885–888.

[27] Z. Wang, C. Chen, and W. Li, "A sequential neural information diffusion model with structure attention," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 1795–1798.

[28] S. Mishra, "Bridging models for popularity prediction on social media," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 810–811.

[29] A. Sankar, X. Zhang, A. Krishnan, and J. Han, "Inf-VAE: A variational autoencoder framework to integrate homophily and influence in diffusion prediction," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 510–518.

[30] J. Wang, H. Zhao, and Z. Liu, "Topic propagation prediction based on dynamic probability model," *IEEE Access*, vol. 7, pp. 58685–58694, 2019.

[31] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1999, pp. 50–57.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

• • •