Research Article

Evolutionary Prediction of Nonstationary Event Popularity Dynamics of Weibo Social Network Using Time-Series Characteristics

Xiaoliang Chen ¹, Xiang Lan ², Jihong Wan ³, Peng Lu ⁴, and Ming Yang ¹

¹School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

²Institute of Statistical Science, Sichuan Provincial Bureau of Statistics, Chengdu 610041, China

³Institute of Artificial Intelligence, School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China

⁴Department of Computer Science and Operations Research, University of Montreal, Montreal QC H3C3J7, Canada

Correspondence should be addressed to Xiang Lan; lanxiang2021@yeah.net

Received 9 February 2021; Accepted 27 April 2021; Published 10 May 2021

Academic Editor: Luca Pancioni

Copyright © 2021 Xiaoliang Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A growing number of web users around the world have started to post their opinions on social media platforms and offer them for share. Building a highly scalable evolution prediction model by means of evolution trend volatility plays a significant role in the operations of enterprise marketing, public opinion supervision, personalized recommendation, and so forth. However, the historical patterns cannot cover the systematical time-series dynamic and volatility features in the prediction problems of a social network. This paper aims to investigate the popularity prediction problem from a time-series perspective utilizing dynamic linear models. First, the stationary and nonstationary time series of Weibo hot events are detected and transformed into time-dependent variables. Second, a systematic general popularity prediction model N-SEP²M is proposed to recognize and predict the nonstationary event propagation of a hot event on the Weibo social network. Third, the explanatory compensation variable social intensity (SI) is introduced to optimize the model N-SEP²M. Experiments on three Weibo hot events with different subject classifications show that our prediction approach is effective for the propagation of hot events with burst traffic.

1. Introduction

Social media platforms, for example, Weibo and Facebook, enable web users to post their views in the virtual communities behind screens. This new type of communication has been well accepted and acclaimed for its apparent advantages of low cost and low user interaction risk. However, the obvious benefits have created potential problems that almost everybody may encounter. An increasing number of social users indulge in passing unverified gossip without doing anything meaningful in the virtual community. Therefore, establishing a highly extensible propagation prediction model has recognized the wisdom of employing social media analysis problems. The related research aims to

provide reasonable technique for avoiding public opinion outbreaks and maintaining social stability.

Information popularity prediction in social networks is usually related to the information diffusion models and information-carrying patterns. The primary idea of popularity analysis is to design or learn a prediction model that can accurately reflect a hot event's information propagation law. On the other hand, the temporal evolution of event popularity can be separated into two types (stationary and nonstationary) according to its trend degree of fluctuation. This work is based on the assumption that a hot event's evolution prediction accuracy is closely correlated to its propagation type in a social network. Some of the current excellent literature on prediction and detection problems in

engineering applications pays particular attention to deep learning [1–3]. It is now well established from a variety of studies (e.g., Lin, 2020 [1]) that a small dataset can train a high-precision model. Some studies of engineering problems benefit a lot by using hybrid deep learning models [2] or deep fusion models [3]. On the other hand, dynamic linear models (DLM) [4, 5] are quite suitable to be employed for the prediction problem concerning their feature extensibility. DLM has a strong interpretability compared with the deep learning models. Hence, this study seeks to obtain a general popularity prediction framework based on DLM, which will help address the evolutionary prediction problem of information propagation dynamics based on stationary theory and time-series characteristics.

The rest of this paper is organized as follows. Section 2 summarizes the related works. Section 3 systematically reviews the evolution analysis models and assessment methods used in this study, including the state space models, dynamic linear models, Kalman filtering, Kalman smoothness and prediction, and maximum likelihood estimation. Section 4 details the proposed information popularity prediction models based on DLM. Section 5 presents the experimental results. Finally, we conclude this work in Section 6.

2. Related Work

Social networks play an essential role in our daily life. People join multiple social network platforms, for example, Facebook and Twitter, to enjoy different services. Information propagation through online social networks has proved to be a powerful tool in many situations. The review in [6] on information propagation has highlighted several application disciplines ranging from biology to social sciences, mathematics, physics, and computer science [7]. The classical application in this field is the virus spread analysis in ecology [8], biology [9], and marketing [10].

Previous studies of information propagation prediction in social networks refer to one of the following three tasks: predicting information popularity [11–14], forecasting user influence [15–18], and divining information diffusion paths (links) [19–22]. Some of the literature focuses on the user influence in the social analysis [15, 17]. Some significant studies are concerned with link prediction to reveal the evolution of real social networks [19, 20]. Much of the current literature on the propagation prediction pays particular attention to information popularity, since it can clearly and intuitively reveal the genuine impact of a hot event by employing statistics. For example, the popularity of video content on social network platforms TikTok and YouTube is often regulated by the statistics such as views, followers, favorites, shares, and downloads [23, 24]. The popularity prediction focuses on the whole trend of information diffusion, for example, the propagation ranges [25] and lifespans [26, 27], to provide the valuable decision-making supports for network public opinion monitoring and guidance. This study sets out to propose the universal scheme of the information popularity prediction problems with time-series characteristics.

Weibo 1 is the unique microblog platform of China which can satisfy regular users' demands to send messages across the country. Any public opinion that might spread in society would certainly rely on the Weibo platform in China. There will always be hundreds of millions of users forwarding their posts to Weibo from the portal, forums, moments, and other media if the content is novel or valuable. Weibo social network is popularly accepted as a monitoring center for public opinion in the era of big data. This study takes Weibo as the carrier to deal with popularity prediction considering its influence on all social networks in China.

The popularity prediction in social networks is a typical time-series problem, in which solutions can provide useful decision supports for avoiding negative propaganda, rumor, hotspot disposal, and so forth. The current research on Weibo popularity prediction is mainly based on the methods of epidemic models [28, 29], classification models [30, 31], and regression models [32, 33]. Such approaches highlight the requirement for time-series analysis of social networks. Time-series analysis can help researchers realize the random mechanism of generating time feature sequences, set up the data generation model, and predict the future possible values of time series. However, a systematic understanding and process of how to analyze time-series problems in the evolution prediction of a social network is still lacking.

Traditional predictive models are usually related to information cascade theory [6], divided into graph-based and non-graph-based ones. The former explore the dynamics of information propagation by individual nodes starting from an initial set of nodes and spreading through the network based on a cascade model, for example, linear threshold models [34] and independent cascade models [35]. The latter mathematically study diffusion using population-based dynamics, for example, SIR models [36]. Some recent attention has focused on the application of incorporating both traditional prediction models and time-series factors [37, 38]. Wu et al. [39] have shown an example to apply the time series of dynamic data in the task of popularity analysis. This method focuses on the regression modeling of time series corresponding to the propagation of the user-generated content. The investigation pattern of user-generated content popularity over time is supported by Hu et al. [40], who utilize regression models and three time-series features to recognize content changes. Matsubara et al. [41] accompanied additional examination of popularity prediction by applying the SpikeM pattern to fit the above-reported time-series models. The published studies have focused on using a temporal feature as an analytical tool rather than concentrating on the model theory of time series in social networks. Some tiny differences that appeared in the popularity of prediction research may lead to consequential large variations in approach.

Some investigators [38, 42] have examined the popularity of the trend analysis in the prediction problem of social networks. Manshad et al. [38] suggested a new time-series trend prediction method based on irregular cellular learning automaton and evolutionary computation. Figueiredo et al. [42] used YouTube videos as samples to extract popular

trends from historical uploaded video objects, combining the new time-series classification algorithm (TrendLearner) with the target features to predict the new target trend.

Popularity trend fluctuation in a social network can be directly exploited for the evolution prediction problem. Some studies deal with the phenomenon of trend fluctuation but lack the investigation of the volatility role in social networks' prediction problems. Wang et al. [37] introduced a time-series prediction method based on complex network theory, which maps the time series to a data network and extracts fluctuation sequence features based on its network topology. The data fluctuation term is recommended as an optimal proposal to the prediction problem for the first time. Hu et al. [43] proposed a time-series feature space containing the average of the prevalence, trend, and period to capture viral hot topics' epidemic behavior. The study observed a high degree of similarity between the short-term trend fluctuations of these hot topics.

Constructing a general evolutionary prediction model with time series is considered one of the most critical social network analysis tasks. Understanding the evolution trend volatility and offering some vital insights into its applications in the popularity prediction problem benefit a lot in the time-series analysis of social networks.

This paper addresses studying the popularity prediction problem from a time-series perspective by means of dynamic linear models. A systematic general popularity prediction model N-SEP²M is proposed to recognize and predict the nonstationary event propagation of a hot event on the Weibo social network. First of all, the popularity evolution of a social network can be distinctly divided into two patterns, stationary (Figure 1(a)) and nonstationary (Figure 1(b)), according to the apparent difference in volatility. The institution concept of the volatility of a time feature sequence shows that the quantitative fluctuation term has a high correlation in the accuracy of event propagation prediction. Our idea requires creating separate prediction models for the stationary and nonstationary events based on volatility calculation.

On top of that, the explanatory compensation variables are introduced into the model N-SEP²M, optimizing dynamic time-series prediction models of the Weibo hotspot event propagation. We summarize the main contributions as follows:

- (i) One of the more significant contributions to emerge from this study is that we recommend a general time-series modeling method of popularity prediction by establishing dynamic linear models based on stationary and nonstationary time-series evolution characteristics in social networks.
- (ii) The benefit of N-SEP²M is that the above social network prediction problems' feature parameters can be updated simply by adding matrix rows and columns of the model, thus avoiding the negative influence on model design and model adaptability for the variations of the target prediction task.
- (iii) This study evaluates N-SEP²M on three real-world hot events of the Weibo social network. The results

show the superior performance by using a proposed explanatory compensation variable, social intensity (SI). The accuracy of the N-SEP²M model can be significantly improved by adding compensation parameters.

3. Preliminaries

This section systematically reviews the evolution analysis technologies based on model theory, aiming to analyze evolution prediction of information diffusion dynamics in social networks. Characterization of time series that appeared in social networks is essential for our increased perception of network information diffusion rules. Simultaneously, we require learning the time-series law of social networks so that a high-precision information diffusion prediction model can be established. Time series refers to a number sequence arranged by successive observations of the same phenomenon at separate times. Statistically, a time series is the realization of a stochastic process, which can be divided into stationary and nonstationary according to its statistical characteristics. However, the series that appear in social networks are, for the most part, nonstationary. The following part of this section describes the preliminaries to establish a popularity evolutionary prediction model of information diffusion dynamics based on time-series characteristics.

3.1. State Space Model. The state space model (SSM) is a dynamic time-domain model with time-dependent variables. A growing number of studies around the world have started to apply SSM in social and economic analysis. One purpose of SSM was to process time sequences of several variables into vector time series, transforming the information cascade problem of the social network into a vector sequence analysis problem. Specifically, an SSM is a dynamic system that evolves over time, which is determined by two different time series. One is called the state sequence, denoted by $\{\theta_t\}$, that is, $\{\theta(t)\}$, with respect to the discrete time variable $t \in N$. Any state belonging to sequence $\{\theta_t\}$ in a system is hidden and unobservable, since the system is inevitably subject to external interference. The other is counterpart observable sequence, denoted by $\{y_t\}$, $t \in N$. The SSM connects the two series by introducing an iterative state equation and an observation equation. The former depicts the transition relation of any adjacent states between the current state θ_{t-1} and the next moment θ_t . The latter indicates the internal relationship between observation and state sequences. An SSM follows two basic assumptions:

- (i) Markov hypothesis: the state sequence $\{\theta_t\}$ constitutes a Markov chain, and the current state is only related to the previous moment.
- (ii) Conditional independence assumption: under the condition of θ_t , the observed values of y_t are independent of each other. In other words, y_t depends only on θ_t .

FIGURE 1: Time-series schematic diagram. The evolution of the stationary time series is usually around a fixed interval, while the evolution interval of the nonstationary time series has apparent variation. (a) Stationary time series. (b) Nonstationary time series.

Based on the above assumptions, the primary form of an SSM can be formalized as the state equation (1) and the observation equation (2):

$$\theta_t = g_t(\theta_{t-1}, \omega_t), \quad (1)$$

$$y_t = h_t(\theta_t, \nu_t), \quad (2)$$

where the mathematical notations ω_t and ν_t represent system noise and measurement noise, respectively. The two noises are typically considered to be the mutually independent Gaussian distributions. The function g_t depicts the transfer relationship of the state variables θ_t . Both g_t and θ_t may be linear or nonlinear. According to the different function forms, the SSM can be roughly divided into three categories: linear, nonlinear, and mixed linear/nonlinear. The complexity increases from the linear to the nonlinear models. Function h_t is used to measure the mapping relationship between the state variables in sequence $\{\theta_t\}$ and the observation variables in sequence $\{y_t\}$.

Constructing an SSM for a specific time-series system is to estimate the state sequences that may occur in the system life cycle. The state estimation problem is a dynamic estimation problem, which can be divided into three types: smoothing, filtering, and prediction. Among them, filtering is the core. The process of state estimation is shown in Figure 2.

- (i) A state estimation problem is called smoothing if we utilize the real-time information up to the current moment t to estimate and trace back the past states. In other words, the noise-removed state values within the total time T , that is, $\forall t < T$, are the target to be traced and identified by investigating the observation sequence $y_1, y_2, \dots, y_t$.
- (ii) A state estimation problem is addressed as filtering if both the real-time system information up to the current moment t and the observed value y_t are used to estimate and revise the current state θ_t .
- (iii) A state estimation problem is designated as prediction if real-time information up to the current

moment t is adopted to predict the future states. The problem estimates the state values at time $t + 1, t + 2, \dots, t + k$, respectively, according to the known observation sequence $y_1, y_2, \dots, y_t$.

In summary, the SSM can be seen as a theoretical algorithm framework, which provides a flexible method for time-series analysis by means of state vectors. The system state prediction based on the SSM is convenient for the analyst employing statistical theory to test the model. Existing research on system state prediction recognizes the critical role played by the SSM. Many economic and financial time-series models can be represented in the form of SSM, such as the autoregressive integrated moving average (ARIMA) model [44], dynamic linear model (DLM) [4, 5], and stochastic volatility model (SVM) [45]. This paper innovatively proposes a general algorithm framework for the time-series analysis of information propagation in social networks by using a DLM. The approach (Section 3.2) introduced for this study is one of a well-designed DLM. An empirical analysis of information dissemination prediction is presented in Section 4 by using the Weibo social network data and DLMs.

3.2. Dynamic Linear Model. The DLM is presented as a particular case of a general SSM with Gaussian and linear characteristics. The estimation and forecasting tasks can be obtained recursively by the well-known Kalman filter of DLM which is the most prevalent and widely accepted approach in analyzing state space problems. The information cascade prediction of social networks can therefore be simplified as a state iteration problem. The characteristic of a DLM is that all model variables obey Gaussian distribution and satisfy linear relationships. A DLM can be represented by the following mathematical form:

$$\theta_t = G_t \theta_{t-1} + \omega_t, \quad \omega_t \sim N_p(0, W_t), \quad (3)$$

$$y_t = F_t \theta_t + \nu_t, \quad \nu_t \sim N_m S(0, V_t). \quad (4)$$

FIGURE 2: The process of state estimation. The initial system state θ_0 and the observed states $\{y_i\}$ from time point $i = 1$ to point $i = t$ are known. The system state θ_{t+1} at time $t + 1$ is required to be predicted by smoothing and filtering operations.

Equations (3) and (4) are the state equation and the observation equation of the DLM, respectively. The uppercase parameters G_t and F_t denote the state transition matrix and measurement matrix, respectively. The parameters ω_t and ν_t are the system noise and measurement noise, respectively. Two kinds of noises are independent of each other, which follow the Gaussian distribution.

Let $\varphi = \{y_t, \theta_t, G_t, F_t, W_t, V_t\}$ denote the parameter set related to a DLM. Then, all parameters in set φ except the observed variables y_t may be unknowns when a DLM is given a practical application task. The dynamic linear model can be used to process parameter estimation tasks. The problem can be divided into three categories according to different parameter estimation targets required in set φ :

- (i) Given the state transition matrix G_t , measurement matrix F_t , system noise W_t , and measurement noise V_t , the estimated target parameter is the state θ_t . This kind of problem is a typical state estimation problem. Then, we generally use the Kalman filtering, smoothing, and prediction algorithms.
- (ii) Given the state transition matrix G_t and measurement matrix F_t , the target is to estimate noise items W_t and V_t . Then, the maximum likelihood estimation approach has obvious advantages.
- (iii) When the parameters θ_t, G_t, F_t, W_t , and V_t are unknown and we require estimating the parameters in them, the Markov Chain Monte Carlo (MCMC) method based on Markov process theory is the potential choice.

The information diffusion evolution prediction of social networks mainly involves the first two categories. The following is a brief description of the Kalman filtering, smoothing, prediction, and the maximum likelihood estimation method applied in this study.

3.2.1. Kalman Filtering. As mentioned above, the state estimation can be divided into prediction, filtering, and smoothing according to the information obtained from the system. Among them, filtering is the core of state estimation. In the filtering problem, the data are supposed to arrive

sequentially in time. We require a procedure to estimate the state vector's current value, based on the observations up to time t , and to update our estimates and forecasts as new data become available at time $t + 1$. The filter process provides the available specifications for updating the current inference on the state vector as new data. We first give the derivation and procedure of the Kalman filtering algorithm.

The Kalman filtering algorithm [46, 47] has been developed to recursively solve the linear filtering problem of discrete data by using the Bayes theorem [48]. The primary task is to predict the prior distribution of the state time, that is, the probability density at the time t , based on the state of the system at time $t - 1$. Then, the algorithm corrects the prior distribution using the Bayes theorem after the likelihood factor, that is, system observation value at time t , is obtained. The result probability density is recognized as the posterior distribution of the state at time t . Therefore, the state estimation of each system moment can be regarded as a normative prediction/correction process. Let the abbreviated expression $y_{1:t}$ with $t \leq T$ denote the observation sequence $\{y_1, y_2, \dots, y_t\}$ and let symbol $\mathbf{p}$ represent a conditional distribution. We formalize the Kalman filter process below to facilitate the introduction of filtering problems of social network information diffusion.

$$\begin{aligned}
 \mathbf{p}(\theta_t | y_{1:t}) &= \frac{\mathbf{p}(y_{1:t} | \theta_t) \mathbf{p}(\theta_t)}{\mathbf{p}(y_{1:t})} \\
 &= \frac{\mathbf{p}(y_t, y_{1:t-1} | \theta_t) \mathbf{p}(\theta_t)}{\mathbf{p}(y_{1:t})} \\
 &= \frac{\mathbf{p}(y_t | y_{1:t-1}, \theta_t) \mathbf{p}(y_{1:t-1} | \theta_t) \mathbf{p}(\theta_t)}{\mathbf{p}(y_t | y_{1:t-1}) \mathbf{p}(y_{1:t-1})} \\
 &= \frac{\mathbf{p}(y_t | y_{1:t-1}, \theta_t) \mathbf{p}(\theta_t | y_{1:t-1})}{\mathbf{p}(y_t | y_{1:t-1})} \\
 &= \frac{\mathbf{p}(y_t | \theta_t) \mathbf{p}(\theta_t | y_{1:t-1})}{\mathbf{p}(y_t | y_{1:t-1})}.
 \end{aligned} \tag{5}$$

The conditional distribution equation (5) describes the probability distribution of the discrete random variable (state value θ_t) under the condition when another discrete random variable (observation sequence $y_{1:t}$) obtains possible fixed values. Both variables θ_t and $y_{1:t}$ follow Gaussian distributions in a DLM model. All of the distributions represented as the form $\mathbf{p}(\cdot)$ mentioned above can be uniquely determined by the expectation and covariances.

The primary operations of the Kalman filtering are predicted, which aim to ascertain the prior distribution and normalization factor of the system states. First of all, we calculate the predicted mean α_t and covariance matrix R_t of state variables at any time t :

$$\alpha_t = E(\theta_t | y_{1:t-1}) = G_t m_{t-1}, \tag{6}$$

$$R_t = \text{Var}(\theta_t | y_{1:t-1}) = G_t G_{t-1}' G_t + W_t, \tag{7}$$

where G_t and G_t' are the state transition matrix and its transpose. Variable notation $m_t = \sum_{k=0}^{k=t} \theta_k \mathbf{p}(\theta_k)$ with respect to $0 \leq t \leq T$ denotes the state mean from θ_0 to θ_t .

On top of that, we estimate the predicted mean f_t and covariance matrix Q_t of the observed variable at time t :

$$f_t = E(y_t | y_{1:t-1}) = F_t \alpha_t, \quad (8)$$

$$Q_t = \text{Var}(y_t | y_{1:t-1}) = F_t R_{t-1}' F_t + V_t, \quad (9)$$

where F_t and F_t' are the measurement matrix and its transpose. V_t denotes the measurement noise. α_t and R_{t-1} are the predicted mean and covariance matrix calculated by equations (6) and (7), respectively.

The correction operations emphasized in the Kalman filtering were based on the parameter values obtained by the prediction operations, the objective of which is to calculate the filtering mean m_t and covariance matrix C_t of the state variable at time t :

$$m_t = E(\theta_t | y_{1:t}) = \alpha_t + R_t' F_t Q_t^{-1} (y_t - f_t), \quad (10)$$

$$C_t = \text{Var}(\theta_t | y_{1:t}) = R_t - R_t' F_t Q_t^{-1} F_t R_t, \quad (11)$$

where Q_t^{-1} is the inverse of covariance matrix Q_t . The detailed procedure of Kalman filtering algorithm is given in Algorithm 1.

3.2.2. Kalman Smoothness and Prediction. Smoothing estimation in time-series problems is mainly used for retrospective analysis of observation sequences to explore potential phenomena or laws underlying the observation values. In the case of many practical problems, the system's a priori parameters cannot be calculated directly, which need to be estimated. For example, in economic research, researchers may necessitate to apply a country's recent gross domestic product to understand the socioeconomic behavior of a country's systems in the past. The forward-backward smoothing algorithm can make such an estimation based on an observation sequence. For each state in a discrete system, it calculates both the "forward" probability of reaching the state and the "backward" probability of generating the model's final state. This study applies the underlying gradient descent idea in the forward-backward smoothing algorithm to modify the past state observation values. Filtering distribution is necessary for the forward-backward smoothing. The derivation process of the filtering distribution is as follows:

$$\begin{aligned} \mathbf{p}(\theta_t | y_{1:T}) &= \int \mathbf{p}(\theta_t, \theta_{t+1} | y_{1:T}) d\theta_{t+1} \\ &= \int \mathbf{p}(\theta_{t+1} | y_{1:T}) \mathbf{p}(\theta_t | \theta_{t+1}, y_{1:T}) d\theta_{t+1} \\ &= \int \mathbf{p}(\theta_{t+1} | y_{1:T}) \frac{\mathbf{p}(\theta_{t+1} | \theta_t) \mathbf{p}(\theta_t | y_{1:t})}{\mathbf{p}(\theta_{t+1} | y_{1:t})} d\theta_{t+1} \\ &= \mathbf{p}(\theta_t | y_{1:t}) \int \mathbf{p}(\theta_{t+1} | \theta_t) \frac{\mathbf{p}(\theta_{t+1} | y_{1:T})}{\mathbf{p}(\theta_{t+1} | y_{1:t})} d\theta_{t+1}. \end{aligned} \quad (12)$$

According to equation (12), the entire smoothing process is reverse propelling by time, which employs the predicted distribution $\mathbf{p}(\theta_{t+1} | y_{1:t})$, the filtering distribution $\mathbf{p}(\theta_t | y_{1:t})$ of the state, and the smooth distribution $\mathbf{p}(\theta_t | y_{1:T})$ of the next moment. Through forward filtering, the filter density distribution $\mathbf{p}(\theta_t | y_{1:t})$ and the predicted density distribution $\mathbf{p}(\theta_{t+1} | y_{1:t})$ at time t can be obtained. Then, we combine them with the smooth density distribution $\mathbf{p}(\theta_{t+1} | y_{1:T})$ at time $t+1$. The objective smooth density distribution $\mathbf{p}(\theta_t | y_{1:T})$ at time t can be deduced backward.

This study takes the output, that is, the state mean m_t and covariance C_t , of the Kalman filter algorithm (Algorithm 1) as the input of forward-backward smoothing and prediction algorithm. The aim is to obtain their estimated values that are similarly denoted by m_t^e and C_t^e , respectively. The algorithm follows a binary classification process. First, the smooth distribution is equal to the filter distribution if the time parameter satisfies $t = T$:

$$\mathbf{p}(\theta_t | y_{1:T}) \sim N(m_t^e = m_t, C_t^e = C_t). \quad (13)$$

The smooth distribution and the normal distribution with parameters m_t and C_t are equivalent under such a branch. Second, if time $t \in \{1: T-1\}$ is an intermediate state time other than the initial time $t = 0$ or the terminal time $t = T$, we calculate the estimated smooth mean m_t^e and the estimated covariance C_t^e for any time $t \in \{1: T-1\}$ according to the following equations:

$$m_t^e = E(\theta_t | y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}^{-1} (m_{t+1}^e - \alpha_{t+1}), \quad (14)$$

$$C_t^e = \text{Var}(\theta_t | y_{1:T}) = C_t + C_t G_{t+1}' R_{t+1}^{-1} (R_{t+1} - C_{t+1}^e) R_{t+1}^{-1} G_{t+1} C_t. \quad (15)$$

Then, the filter distribution $\mathbf{p}(\theta_t | y_{1:T}) \sim N(m_t^e, C_t^e)$ can be obtained successfully.

Under the assumption of the Gaussian distribution related to the DLM models, the execution process of the forward-backward smoothing algorithm can be summarized as Algorithm 2.

As explained in Algorithm 2, it is clear that the filter distribution has been determined under any time t . Subsequently, we can deal with the prediction problems involved in the observed values or state values at any future time point $t+k$ ($k > 0$) by a recursive prediction process. The calculation detail is as follows:

(i) Prediction of state variables:

$$\begin{aligned} \alpha_t(k) &= E(\theta_{t+k} | y_{1:t}) = G_{t+k} \alpha_t(k-1), \\ R_t(k) &= \text{Var}(\theta_{t+k} | y_{1:t}) = G_{t+k} R_{t,k-1} G_{t+k}' + W_{t,k}. \end{aligned} \quad (16)$$

(ii) Prediction of observation variable:

$$\begin{aligned} f_t(k) &= E(y_{t+k} | y_{1:t}) = F_{t+k} \alpha_t(k), \\ Q_t(k) &= \text{Var}(y_{t+k} | y_{1:t}) = F_{t+k} R_t(k) F_{t+k}' + V_{t,k}. \end{aligned} \quad (17)$$

Input: Initial state mean m_0 and initial state covariance matrix C_0 .
Output: State mean m_t and state covariance matrix C_t at each time t .
(1) **for** $\forall t \in [1, 2, \dots, T]$ **do**
(2) Prediction operations:
(3) $\alpha_t = E(\theta_t | y_{1:t-1}) = G_t m_{t-1}$ equation (6)
(4) $R_t = \text{Var}(\theta_t | y_{1:t-1}) = G_t G_{t-1}' G_t' + W_t$ equation (7)
(5) $f_t = E(y_t | y_{1:t-1}) = F_t \alpha_t$ equation (8)
(6) $Q_t = \text{Var}(y_t | y_{1:t-1}) = F_t R_{t-1} F_t' + V_t$ equation (9)
(7) Correction operations:
(8) $m_t = E(\theta_t | y_{1:t}) = \alpha_t + R_t F_t' Q_t^{-1} (y_t - f_t)$ equation (10)
(9) $C_t = \text{Var}(\theta_t | y_{1:t}) = R_t - R_t F_t' Q_t^{-1} F_t R_t$ equation (11)
(10) **end**

ALGORITHM 1: Kalman filtering process.

Input: State filter mean m_t and covariance C_t at time $t \in \{1: T\}$.
Output: State mean estimate m_t^e and covariance estimate C_t^e at time $t \in \{1: T\}$.
(1) **if** $t = T$ **then**
(2) $\mathbf{p}(\theta_t | y_{1:T}) \sim N(m_t^e = m_t, C_t^e = C_t)$ equation (13)
(3) **else if** $t \in \{1: T-1\}$ **then**
(4) **for** $\forall t \in \{1: T-1\}$ **do**
(5) $m_t^e = E(\theta_t | y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}^{-1} (m_{t+1}^e - \alpha_{t+1})$ equation (14)
(6) $C_t^e = \text{Var}(\theta_t | y_{1:T}) = C_t + C_t G_{t+1}' R_{t+1}^{-1} (R_{t+1} - C_{t+1}^e) R_{t+1}^{-1} G_t C_t$ equation (15)
(7) $\mathbf{p}(\theta_t | y_{1:T}) \sim N(m_t^e, C_t^e)$.
(8) **end**
(9) **end**

ALGORITHM 2: Kalman forward-backward smoothing algorithm.

3.2.3. Maximum Likelihood Estimation. Maximum likelihood estimation (MLE) is one of the most important and widely used parameter estimation methods in statistics. The idea is to determine the unknown parameters in a system model by maximizing the probability of observed values. Under the condition that the state transition matrix G_t and the measurement matrix F_t are known from the system environment, the maximum likelihood estimation of noise terms can be derived.

This section uses the simplified expression $y_{1:T}$ to describe the observation data $\{y_1, y_2, \dots, y_T\}$, whose distribution depends on the state noise W and the observation noise V . Let the joint probability density $L(V, W | y_{1:T})$ denote the likelihood function of noise items V and W relative to the observation sequence $y_{1:T}$. We have the following formal description:

$$\begin{aligned} L(V, W | y_{1:T}) &= L(V, W | y_1, y_2, \dots, y_T) \\ &= \prod_{t=1}^T \mathbf{p}(V, W | (y_t | y_{1:t-1})). \end{aligned} \quad (18)$$

The derivation $\mathbf{p}(V, W | (y_t | y_{1:t-1}))$ in equation (18) is the prediction probability density function of the observation sequence $y_{1:t-1}$, which obeys the Gaussian distribution with mean f_t and variance Q_t obtained by the forward

Kalman filter algorithm. To simplify the operation, we convert equation (18) to its log-likelihood form:

$$L(V, W | y_{1:T}) = -\frac{1}{2} \sum_{t=1}^p \log |Q_t| - \frac{1}{2} \sum_{t=1}^p (y_t - f_t)' Q_t^{-1} (y_t - f_t). \quad (19)$$

The purpose of equation (19) is to find the noise terms V and W , which makes the occurrence probability of observation sequence $y_{1:T}$ in the system the greatest. According to the principle of the MLE, parameters V and W can be estimated only by dealing with the optimization task of maximizing equation (19). Finally, we use the quasi-Newton method [49] to solve the parameters.

4. Information Popularity Prediction in Social Networks Based on Dynamic Linear Models

A growing number of information senders and receivers around the world have started to post their speeches, thoughts, and comments online on some new social media platforms such as Weibo, Facebook, Twitter, WeChat, and LinkedIn. These attention-grabbing open media have been well accepted and acclaimed for their apparent advantages: faster transmission speed, being more timely, broader coverage, higher public, and being more interactive. These types of social network platforms will play an increasingly

significant role in our future life, society, business, and so forth.

The graphical properties of the users and the relationships between users in a social platform are called social networks. The information popularity prediction in social networks is a typical time-series problem, the solutions of which can provide useful decision supports for avoiding negative propaganda, rumor, hotspot disposal, and so forth. The issue of time-series analysis in social networks has received considerable attention. However, a systematic understanding and process of how to analyze time-series problems in social networks is still lacking.

This section takes Weibo as an example to discuss how the DLM can be used as a standard analysis framework for the propagation and prediction issues of social networks with time-series characteristics. The study first highlights some important statistics and evolution rules of the Weibo social network, including the number of daily posts, retweets, comments, and how they change with time. Second, we use the theory of the DLM, combined with social data and search data, to decompose the hotspot events' time series and predict their propagation effect on the Weibo network.

In this paper, the time series of information propagation in Weibo is regarded as the superposition of trending, periodic, and random terms.

- (i) The trend term characterizes the objective evolutionary trend of Weibo hotspot events.
- (ii) The periodic term describes the regular changes due to cyclic factors such as the incubation, development, outbreak, mature, recession, and extinction periods of a network public opinion event.
- (iii) The random term involves a sudden change or noise disturbance. A sudden change refers to a change caused by unexpected circumstances such as the hot news retweet or emergent unexpected social events. The noise illustrates the influence of many random factors, such as different attitudes of users who release or forward messages, public opinion reversal caused by facts, and public opinion explosion.

The DLM model can describe the parameters of the information propagation time series listed above. The proposed models are given in Section 4.3.

4.1. Challenges of Information Propagation Prediction. A well-structured propagation prediction model benefits researchers or government departments to find the regularity, suddenness, or reversibility in the process of information propagation. Predicting the spread of a particular news event is helpful to guide the negative public opinion in advance. There are two challenges facing the research on the information propagation model and the prediction of the online social networks that are similar to Weibo:

- (1) Avoidance of low-quality data: Spam users (non-zombies) are flooded in every Weibo platform corner. Many users engaged in Taobao,

microbusiness, and network shopping platforms continue to release advertisements, order display, and purchasing agency information on Weibo. Such jumbled information results in low-quality data. Despite the importance of statistical characteristics such as the number of posts, comments, retweets, and mentions of a hot event on the Weibo social network, there remains a paucity acceptable accuracy of information propagation prediction based on only the specific numbers.

- (2) Data association mining: The essence of prediction analysis is to find the law of the event occurrence from the mass data. The timestamp of a time-series system cannot be used directly for prediction. We still require discovering the underlying factors that influence the time-series state change.

4.2. Solution. This study adopts a divide-and-conquer strategy in response to the issue of low-quality data. The propagation of a hot event on Weibo can be divided into stationary and nonstationary by calculating the event's time-series volatility. First of all, we use DLM models established by using the historical time-series information to deal with stationary event popularity prediction. On top of that, some explanatory variables (hot event characteristic variables) are captured to enhance the DLM models. The improved models are applied to deal with the propagation prediction of nonstationary situations.

Association mining methods in [50] are introduced to find explanatory variables from the Weibo platform by executing information filtering, text classification, data normalization, and so forth. Then, we calculate the correlation coefficient between the event propagation time series and the explanatory variable time series. The resulting highly correlated variables were chosen as the random terms of the time series in the process of event propagation. The complete framework for information popularity prediction of a hot event on the Weibo social network is shown in Figure 3.

4.3. Modeling of Weibo Hot Events Prediction Based on Dynamic Linear Models

Definition 1. A DLM is called a stationary event propagation prediction model (SEP²M) if it has the characteristic of local linear trend and seasonality and possesses the following state equation form:

$$y_t = \mu_t + m_t^e + \nu_t \nu_t \sim N(0, V), \quad (20)$$

where

$$\begin{aligned} \mu_t &= \mu_{t-1} + \delta_{t-1} + \omega_{\mu t}, \omega_{\mu t} \sim N(0, \sigma_\mu^2), \\ \delta_t &= \delta_{t-1} + \omega_{\delta t}, \omega_{\delta t} \sim N(0, \sigma_\delta^2), \\ m_t^e &= -m_{t-1}^e - \dots - m_{t-\eta}^e + \omega_{m_t^e}, \omega_{m_t^e} \sim N(0, \sigma_m^2). \end{aligned} \quad (21)$$

Definition 1 provides a general model to deal with the popularity prediction problems with stationary event

FIGURE 3: A popularity prediction framework for information propagation of a hot event on the Weibo social network. First, the explanatory (time) variables of an event were determined by the data preprocessing, and the key variables affecting the event propagation were identified by the correlation analysis. Second, the framework recognizes the propagation type (stationary or nonstationary) of the event. Third, a DLM is established based on the variables and the event propagation type. Consequently, time-series analysis can be conveniently achieved based on the DLM.

behavior in social networks, where μ_t and δ_t are the level trend and the inclination trend at time t , respectively. m_t^e denotes the state mean estimate of a periodic item m_t , which can be obtained by Algorithm 2. The period length is represented by the Greek symbol η . For example, if the model takes a week (seven days) as a cycle, we have $\eta = 6$. The information propagation of the seventh day is predicted using historical data from the previous six days. The tail terms ν_t , $\omega_{\mu t}$, $\omega_{\delta t}$, and $\omega_{m_t^e}$ of the equations are the measurement noise, level trend noise, inclination trend noise, and mean estimate noise, respectively.

The model representation of SEP²M in matrix form serves as an effective way to perform programming and the subsequent data experiments. Hence, we convert the SEP²M described in Definition 1 into the following matrix form:

(i) The observation equation:

$$y_t = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} \mu_t \\ \delta_t \\ m_t^e \\ m_{t-1}^e \\ m_{t-2}^e \\ \vdots \\ m_{t-\eta-1}^e \end{bmatrix} + \nu_t. \quad (22)$$

(ii) The state equation:

$$\begin{bmatrix} \mu_t \\ \delta_t \\ m_t^e \\ m_{t-1}^e \\ m_{t-2}^e \\ \vdots \\ m_{t-\eta-1}^e \end{bmatrix} = \begin{bmatrix} G_\mu & 0 \\ 0 & G_m \end{bmatrix} \cdot \begin{bmatrix} \mu_t \\ \delta_t \\ m_{t-1}^e \\ m_{t-1}^e \\ \vdots \\ m_{t-\eta-1}^e \\ m_{t-\eta}^e \end{bmatrix} + \begin{bmatrix} W_{\mu_t} \\ W_{\delta_t} \\ W_{m_t^e} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (23)$$

where we have the two coefficient matrices:

$$G_\mu = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad G_m = \begin{bmatrix} -1 & \dots & -1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}. \quad (24)$$

Definition 2. A DLM is called a nonstationary event propagation prediction model (N-SEP²M) if it has the characteristic of local linear trend, seasonality, regression, and possesses the following state equation form:

$$y_t = \mu_t + m_t^e + \beta^T X_t + \nu_t, \quad \nu_t \sim N(0, V), \quad (25)$$

where

$$\begin{aligned}
\mu_t &= \mu_{t-1} + \delta_{t-1} + \omega_{\mu t}, \quad \omega_{\mu t} \sim N(0, \sigma_\mu^2), \\
\delta_t &= \delta_{t-1} + \omega_{\delta t}, \quad \omega_{\delta t} \sim N(0, \sigma_\delta^2), \\
m_t^e &= -m_{t-1}^e - \dots - m_{t-\eta}^e + \omega_{m_t^e}, \quad \omega_{m_t^e} \sim N(0, \sigma_m^2), \\
\beta_t &= \{\beta_{1t}, \beta_{2t}, \dots, \beta_{nt}\}, \\
X_t &= \{x_{1t}, x_{2t}, \dots, x_{nt}\}.
\end{aligned} \quad (26)$$

Definition 2 presents a model suitable for analyzing the propagation of nonstationary events in social networks, where μ_t , δ_t , and m_t^e are similarly the level trend, the inclination trend, and periodic item estimate at time t , respectively. Parameters ν_t , $\omega_{\mu t}$, $\omega_{\delta t}$, and $\omega_{m_t^e}$ are the noises related to the system, level trend, inclination trend, and periodic item mean calculation, respectively. X_t represents explanatory variables. The explanatory variable coefficient is expressed as the symbol β^T . The compound expression $\beta^T X_t$ denotes a random term in a nonstationary event that suffers a sudden occurrence or reversal. We convert the model N-SEP²M into the following matrix form:

(i) The observation equation:

$$y_t = [1, 0, 1, 0, 0, 0, 0, 0, 1, \dots, 1] \cdot \begin{bmatrix} \mu_t \\ \delta_t \\ m_t^e \\ m_{t-1}^e \\ \vdots \\ m_{t-\eta-1}^e \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \nu_t. \quad (27)$$

(ii) The state equation:

$$\begin{bmatrix} \mu_t \\ \delta_t \\ m_t^e \\ m_{t-1}^e \\ \vdots \\ m_{t-\eta-1}^e \\ \beta_{1,t} \\ \beta_{2,t} \\ \vdots \\ \beta_{n,t} \end{bmatrix} = \begin{bmatrix} G_\mu & 0 & 0 \\ 0 & G_m & 0 \\ 0 & 0 & G_x \end{bmatrix} \cdot \begin{bmatrix} \mu_t \\ \delta_t \\ m_{t-1}^e \\ m_{t-2}^e \\ \vdots \\ m_{t-\eta}^e \\ \beta_{1,t-1} \\ \beta_{2,t-1} \\ \vdots \\ \beta_{n,t-1} \end{bmatrix} + \begin{bmatrix} W_{\mu_t} \\ W_{\delta_t} \\ W_{m_t^e} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (28)$$

where the details of coefficient matrices are as follows:

$$\begin{aligned}
G_\mu &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \\
G_m &= \begin{bmatrix} -1 & \dots & -1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \\
G_x &= \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.
\end{aligned} \quad (29)$$

5. Experiments

This section discusses the datasets, stationary and nonstationary identification, and the model evolution analysis of three hot events that appeared on the Weibo social network.

5.1. Event Datasets and Propagation Network Setting. To estimate the effectiveness of the proposed models for popularity prediction, the survey collected the datasets from the Chinese popular social platform Sina Weibo 2, including three Weibo hot public opinion events with the three content categories of social, entertainment, and international news from 2015 to 2017. A brief description of the three events is presented as follows:

(i) *Chengdu female driver incident (Chengdu-Driver).*

On the afternoon of May 3, 2015, a beating incident occurred near the Jiaozi overpass on Chengdu Third Ring Road. Ms. Lu was forced to stop by Zhang at the Jiaozi interchange for changing lanes and was later beaten and injured. The incident was broadcast on Weibo. People were first shocked by the ferocity of Zhang. However, public opinion turned to blame the female driver for driving too unruly and dangerously when the event video came out. The propagation of the event has formed the typical public opinion reversal effect.

(ii) *Celebrity Bai Baihe event (Baihe-Cheating).* The cheating of a female star, Bai Baihe, led to a backlash from her fans and formed a long-term discussion on the behavior of the entertainer on Weibo.

(iii) *THAAD incident.* South Korea's Lotte Group decided to transfer the Seongju Golf Course to the South Korean defense ministry to deploy the THAAD system. The event triggered an immediate outcry in China against Rakuten. Many businesses and people are boycotting Lotte Group's operations in China. With THAAD's deployment in South Korea, the anti-South-Korean sentiment in China, especially Lotte Group, has also been increasing. The incident eventually led to the closure of 55 Lotte Mart stores in China.

The datasets “Chengdu female driver was beaten.csv,” “Baihe Bai derailed.csv,” and “THADD incident.csv” refer to the social event (Chengdu-Driver), an entertainment event (Baihe), and an international news event (THAAD), respectively.

The original hot event data is prepared by crawling user information pages involved in the three event topics, including the post contents, user attributes, follower relations, the number of comments and thumb up, post time, and forwarding (propagation) links. This study implements the data crawling and the proposed prediction models by the programming language Python version 3.7 and runs the codes on a Linux Server with Intel(R) Xeon(R) CPU (E5-2620 v4) and GeForce TITAN X GPU (12 GB memory). The number of valid entries is 95%, 96%, and 98% of the total results returning from our crawler system, respectively. The details about the event datasets are shown in Table 1. The trends for the development of the three events’ datasets are shown in Figure 4.

Our experiments will evaluate and predict the popularity of the three hot events on the Weibo social network. The best environment setting is to get the full structure of Weibo as a map of the spread. However, Weibo allows only a small part of follower relations to be returned by crawlers. Hence, the relations that come from traditional Weibo crawling methods are quite incomplete. On the other hand, the size of Weibo network is enormous. The empirical treatment is to extract a subnet with common characteristics from the original Weibo network. We repeatedly remove the nodes with the degrees less than 2 or more than 1000. Then, the community discovery algorithm reported in [51] is performed to find the subnet that preserves the neighbors and relations of the event-related users. The subnet has the typical social network characteristics, including the approximate power-law degree distribution (Figure 5) and the high aggregation coefficient.

5.2. Predictive Evaluation

5.2.1. Stationary Sequence Identification and Social Intensity. Historically, the statistical concept “volatility” has been used to describe the degree of return fluctuation on the investment in financial markets. Using this concept to distinguish the stationary and nonstationary events is recognized as an essential experimental step. The experiment requires to be explicit about exactly how to obtain the “volatility” of a hot event on Weibo social network. The calculation process is defined as follows.

- (1) Calculate the percentage return:

$$y_t = \frac{(X_{t+1} - X_t)}{X_t}. \quad (30)$$

- (2) Calculate the volatility (standard deviation) of sequence $\{y_t\}$:

$$V = \sqrt{\frac{\sum_{t=1}^n (y_t - \bar{y})^2}{n}}, \quad (31)$$

where the sequence $\{X_t\}$ represents the values of an explanatory variable at different time t , which is also introduced in equation (25) at Section 4.3. $\bar{y}$ is the average value of the sequence $\{y_t\}$. The forwarding behavior of users on Weibo has been playing an increasingly important role in helping researchers get the active degree of event discussion. The forwarding time series of an event is the concrete expression of the law about event propagation. Hence, our experiments consider the forwarding number of an event on Weibo at time t as the explanatory variable X_t . Then, the sequence $\{y_t\}$ of percentage return and the subsequent volatility can be calculated. We follow the subsequent steps to identify the stationary/nonstationary state of an event:

- (1) The life cycle of an event is divided into time-series intervals in days.
- (2) Count the forwarding numbers from 8:00 to 24:00 at the interval of 1 hour within a day (obtaining sequence $\{X_t\}$ for each day).
- (3) The percentage return is calculated to obtain the sequence y_t according to equation (27).
- (4) Calculate the day-to-day volatility, denoted by V_d according to equation (28).
- (5) The possible value of volatility is divided into several (set as nine) intervals. The forwarding numbers of the event in different volatility intervals are calculated by simple numerical statistics.
- (6) Estimate the ratio of the forwarding number to the total number of retweets with different volatility intervals.
- (7) Calculate the sum of the forwarding ratios with volatility values greater than 1. If the parameter is more than 75%, the event is decided to be nonstationary; otherwise, it is stationary.

The threshold of 75% is used to distinguish the application scopes of the two models SEP²M and NSEP²M when we need to represent, analyze, and predict the process of event popularity evolution on the Weibo social network. The annual report of hot searches on Weibo 3 shows that about 90% of the hot events in Weibo are nonstationary. The volatility distribution of an example using the Chengdu-Driver dataset is shown in Figure 6.

Figure 6 presents the statistical relationship between the forwarding numbers and the volatility intervals for the event Chengdu-Driver. The forwarding longitudinal series is a measure of the depth at which an event post is forwarded on the Weibo social network. Series 1–4 shown in Figures 6(a) and 6(b) illustrate that the statistical results of forwarding numbers contain a minimum of one (resp., three) forwarding action and a maximum of four forwarding longitudinal actions. We plot the data distribution of volatility intervals in a descending order. The right label list of proportions shows the ratio of retweets with different volatility intervals to the total forwarding number. The cumulative line (orange color line) on the second axis identifies the cumulative value from the fluctuating high to the low point

TABLE 1: Details about the three datasets of hot public opinion events.

Datasets	Classification	Users	Followers	Forwarding statistic
Chengdu-Driver	Social	7259	135455	159711
Baihe	Entertainment	1533	1395880	536674
THAAD	International news	5532	1891069	2130227

FIGURE 4: Trends for the development of the three Weibo hot events. The social event is more likely to receive social attention according to an earlier increase of forwarding behaviors, followed by entertainment and international news.

as a percentage of the total retweets based on the forwarding numbers.

The red point shown in Figure 6 describes the parameter value to identify the stationary state of the event. As a result, the propagation of social event Chengdu-Driver on the Weibo network is stationary, since the expression $0.78 > 0.75$ is true. If the experiment ignores the forwarding data of series 1 and 2, the decision point calculation value shown in Figure 6(b) is 0.88. A decision point larger than 0.78 indicates that the vertical propagation of depth causes the event's principal fluctuation. Similar analyses are applied in the events Baihe and THAAD. We can finally obtain the decision points 0.83 and 0.91, which are presented in Figures 7(a) and 7(b), respectively. What is interesting about the data in these three points (0.78, 0.83, and 0.91) is that the international news event THAAD had the most significant fluctuations in the three Weibo events. The nonstationary state implies a large amount of discussion in a short period.

5.2.2. Correlation Analysis of Explanatory Variables. Having explained what is meant by explanatory variables, we will now discuss how to find the best explanatory variable. A simple linear regression model is used to establish the quantitative relationships between event data and explanatory variables to determine their correlation scores. This study selects the best explanatory variable by comparing the Pearson Correlation Coefficients between all variables and the forwarding numbers at time series.

We refer to the percentage of forwarding numbers, comment numbers, and thumb-up numbers of an event at time t as the social intensity, comment intensity, and thumb-up intensity, respectively. The main temporal explanatory variables involved in the analysis of the three experimental datasets included the following: comment number (CN), thumb-up number (TUN), social intensity (SI), comment intensity (CI), thumb-up intensity (TUI), and average comment length (AveL). All the above explanatory variables are time-dependent. Table 2 shows the correlation scores between the six explanatory variables and the forwarding numbers with respect to the three events. The experiment filtered out the explanatory variables with low correlation and temporal fluctuation. We determine the best one, that is, social intensity, for the three datasets according to the correlation coefficient comparison.

In general, the higher the correlation between the explanatory variable sequence and the target sequence, the better the prediction effect. Figure 8 shows the quantitative relationship between the event popularity evolution and the social intensity of the event Chengdu-Driver, and the dataset has been normalized.

5.3. Experimental Results and Case Analysis. In order to evaluate the effect of the time-series model and the temporal model after the introduction of the explanatory variable SI, we compare the gradient boosting regression tree (GBRT) model based on model combinations. The evaluation

FIGURE 5: The degree distributions of the Weibo subnet used in the experiments.

FIGURE 6: Stationary sequence recognition result on the event Chengdu-Driver: (a) forwarding longitudinal series 1-4; (b) forwarding longitudinal series 3-4. The sum of the forwarding ratios with volatility values greater than 1 equals 0.78 which is shown in (a). The score indicates the nonstationary characteristic of the event. The instability is mainly caused by the forwarding behavior of levels 3 and 4 according to the value 0.88 shown in (b).

FIGURE 7: Stationary sequence recognition results on the events Baihe and THAAD. The degree of nonstationarity of the three events is sorted as THAAD (0.91), Baihe (0.83), and Chengdu-Driver (0.78).

TABLE 2: Correlation analysis results of explanatory variables.

Events	CN	TUN	SI	CI	TUI	AveL
Chengdu-Driver	0.7233	0.6725	0.8624	0.6658	0.5413	0.1442
Baihe	0.7297	0.5425	0.8756	0.6835	0.5511	0.2234
THAAD	0.7784	0.5265	0.9143	0.8011	0.5643	0.1143

FIGURE 8: The quantitative relationship between the event popularity evolution and the social intensity of the event Chengdu-Driver. An apparent positive correlation between popularity and social intensity can be observed.

indexes adopted include the following: the determination coefficient R^2 , the root mean square error $RMSE$, and the accuracy of the predicted absolute error between 20% and 50% (namely, Precision@20 and Precision@50). The results of the comparison of the evaluation indicators of prediction models are shown in Table 3.

The results of this study indicate that when the popularity prediction of a Weibo hot event only depends on the time information, the error between the prediction result

and the real event propagation is usually large. By comparing three hot Weibo events, we can see from Figures 9, 10, and 11 that using highly correlated explanatory variables as the proposed model parameters can improve the accuracy and reliability of event popularity prediction.

We illustrate the experimental results by using event Chengdu-Driver as an example. The blue trend line recorded in Figure 9(a) shows the observation retweets of the social event in a month. The distinct fluctuation of observation to

TABLE 3: Comparison of the evaluation indicators of prediction models.

Prediction model	R^2	Precision@20	Precision@50	RMSE
GBRT	0.7645	41%	76%	103.45
SEP ² M	0.8761	44%	77%	89.63
N -SEP ² M	0.8975	52%	81%	88.74

FIGURE 9: Popularity evolution prediction of social event: “Chengdu female driver was beaten by changing routes.” (a) Prediction accuracy comparison. (b) Cumulative relative error analysis. The predicted result of the NSEPP model with explanatory variables is the closest to the actual evolution. Social intensity can significantly reduce the cumulative error.

FIGURE 10: Popularity evolution prediction of entertainment event: “Baihe Bai derailed.” (a) Prediction accuracy comparison. (b) Cumulative relative error analysis. A similar prediction result is obtained concerning a slightly rising error compared with Figure 9.

emerge from Figure 9(a) was that there exist the bursts of network traffic flow around the third day. The green and orange lines represent the fitting results with or without the

explanatory variable SI, respectively. The purple-dotted line and the red-dotted line are the corresponding prediction results, respectively. Following the time going on, the fitness

FIGURE 11: Popularity evolution of international news event: "THAAD." (a) Prediction accuracy comparison. (b) Cumulative relative error analysis. A similar prediction result is obtained concerning a slightly rising error compared with Figure 9.

FIGURE 12: Time-series decomposition of nonstationary event "THAAD" with burst traffic. (a) The trend μ_t of the burst flow event. (b) The periodicity m_t^e of the burst flow event. (c) The regression results γ_t of the burst flow event.

values for propagating the event Chengdu-Driver which almost coincide with the real trend lines were recorded. A comparison of the two results exposed by the green and orange trend lines in Figure 9(a) reveals that the prediction accuracy of the NSEPP model with explanatory variables is significantly higher than that of models without considering explanatory variables. Furthermore, it can also be seen from the cumulative relative error recorded in Figure 9(b) that the prediction results are more efficient when the explanatory variables are introduced into the N-SEP²M model for the nonstationary event popularity evolution with burst traffic.

The entertainment event Baihe (Figure 10) and the news event THAAD (Figure 11) showed similar accuracy rate advantage of data fitting in the prediction results with explanatory variables. Comparing the model prediction results of fitting with SI and the original fitting values, the model presented in this study has obvious expected returns. In contrast to the prediction result of event Chengdu-Driver, however, a slightly larger error of the events Baihe and THAAD of the fitting value with SI was detected. The green trend line is not entirely close to the observation line in the prediction charts. The difference is due to the underlying quality of the three datasets. Together, these results provide important good insights for hot event propagations into the significance of combining dynamic linear models and highly correlated explanatory variables.

The common feature of the three Weibo hotspot events is that they all have burst traffic characteristics. In order to verify the adaptability of these events to the parameters defined by the prediction models, the event THAAD with the highest volatility was selected to decompose the time-series quantification values. Figure 11(a) represents the observation, fitting, and prediction of THAAD with burst traffic. Figures 12(a), 12(b), and 12(c) represent the trend μ_t , periodicity m_t^e , and regression results y_t of the burst flow event, respectively, reflecting the characteristics of its nonstationary time series. The decomposition results further show that the proposed dynamic linear model N-SEP²M (Def. ??) can accurately deal with the prediction task of hot event propagation on the Weibo social network.

6. Conclusion

This research is undertaken to design a time-series analysis framework and provide a generic solution for the evolutionary prediction of information popularity dynamics in social networks. This study also aims to alleviate the contradiction between stationary and nonstationary event propagation in a standard prediction model. The fact that all hot events in a social network are analyzed in a class of models without considering their volatility may not necessarily serve as an effective way to increase the prediction accuracy. Hence, the framework's investigation divides event propagation into two kinds (stationary and nonstationary) according to event volatility. This study defines two specific DLM models, namely, SEP²M and N-SEP²M, to analyze stationary and nonstationary propagation, respectively.

This paper mainly reveals the superiority of using the N-SEP²M model to predict nonstationary popularity evolution. N-SEP²M is innovatively constructed by introducing the nonstationary feature parameters of local linear trend, seasonality, and regression. Three different parameter estimation methods, Kalman filter, Kalman smoothing, and maximum likelihood, are used to simplify the parameter adjustment process.

Experiments on three popular hot events that appeared on the Weibo social network with different topic categories confirm the effectiveness and superiority of the N-SEP²M model through the comparison of evaluation indexes R^2 , RMSE, *Precision*@20, and *Precision*@50. The experimental propagation trend lines also exhibit that the prediction accuracy of event popularity evolution can be significantly improved by introducing explanatory variables into the N-SEP²M model. The insights of model construction gained from this study may be of benefit in the complex hot event evolution analysis in social networks with time-series characteristics.

Data Availability

The data used to support the findings of this study are publicly available at <https://github.com/YangMin-10/Datasets>.

Conflicts of Interest

The authors declare that they have no conflicts of Interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61902324, 11426179, and 61872298), the Social Science Planning Project of Sichuan Province (no. SC20TJ020), the Science and Technology Program of Sichuan Province (nos. 2021YFQ0008, 2020JDR0067, and 2019GFW131), the Foundation of Cyberspace Security Key Laboratory of Sichuan Higher Education Institutions (no. sjzz2016-73), the Scientific Research Fund of Sichuan Provincial Education Committee (nos. 15ZB0134 and 17ZA0360), and the Open Fund Project of Xihua University (nos. 20170410143123 and szjj2015-059).

References

- [1] Y. Lin, L. C. Li, H. L. Jing, B. Ran, and D. Y. Sun, "Automated traffic incident detection with a smaller dataset based on generative adversarial networks," *Accident Analysis and Prevention*, vol. 144, 2020.
- [2] L. C. Li, Y. Lin, B. W. Du, F. Yang, and B. Ran, "Real-time traffic incident detection based on a hybrid deep learning model," *Transportmetrica A: Transport Science*, 2020.
- [3] L. C. Li, X. Sheng, B. W. Du, Y. G. Wang, and B. Ran, "A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction," *Engineering Applications of Artificial Intelligence*, vol. 93, 2020.
- [4] J. Meng, D.-I. Stroe, M. Ricco, G. Luo, and R. Teodorescu, "A simplified model-based state-of-charge estimation approach

- for lithium-ion battery with dynamic linear model," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 10, pp. 7717–7727, 2019.
- [5] Q. Cheng, J. Xing, W. Yi, Z. Liu, and X. Fu, "Distance-based congestion pricing with day-to-day dynamic traffic flow evolution process," *Discrete Dynamics in Nature and Society*, vol. 2019, Article ID 7438147, 7 pages, 2019.
 - [6] M. Jalili and M. Perc, "Information cascades in complex networks," *Journal of Complex Networks*, vol. 5, pp. 665–693, 2017.
 - [7] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, "Dynamics of information diffusion and its applications on complex networks," *Physics Reports*, vol. 651, pp. 1–34, 2016.
 - [8] V. Colizza, A. Barrat, M. Barthélemy, A. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions," *PLoS Medicine*, vol. 4, no. 1, pp. 95–110, 2007.
 - [9] L. Stone, R. Olinky, and A. Huppert, "Seasonal dynamics of recurrent epidemics," *Nature*, vol. 446, no. 7135, pp. 533–536, 2007.
 - [10] P. Domingos, "Mining social networks for viral marketing," *Journal of Retailing and Consumer Services*, vol. 20, no. 1, pp. 24–28, 2010.
 - [11] H. T. Li, X. Q. Ma, F. Wang, J. C. Liu, and K. Xu, "On popularity prediction of videos shared in online social networks," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pp. 169–178, ACM, New York, NY, USA, 2013.
 - [12] P. Bao, H. W. Shen, J. M. Huang, and X. Q. Cheng, "Popularity prediction in microblogging network: a case study on sina weibo," in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 177–178, ACM, New York, NY, USA, 2013.
 - [13] Q. Kong, W. Mao, and C. Liu, "Popularity prediction based on interactions of online contents," in *Proceedings of the 4th International Conference on Cloud Computing and Intelligence Systems*, pp. 1–5, ACM, Beijing, China, 2016.
 - [14] Q. Cao, H. W. Shen, J. H. Gao, B. Z. Wei, and X. Q. Cheng, "Popularity prediction on social platforms with coupled graph neural networks," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 70–78, ACM, New York, NY, USA, 2020.
 - [15] N. Alduaiji, A. Datta, and J. Li, "Influence propagation model for clique-based community detection in social networks," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 563–575, 2018.
 - [16] M. Ra2l, G. Ortiz, M. Postigo-Boix, L. Jos2s, and M. Mels, "Extent prediction of the information and influence propagation in online social networks," *Computational and Mathematical Organization Theory*, 2020.
 - [17] J. Li, W. Peng, T. Li, T. Sun, Q. Li, and J. Xu, "Social network user influence sense-making and dynamics prediction," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5115–5124, 2014.
 - [18] R. Pagare and A. Khare, "Churn prediction by finding most influential nodes in social network," in *Proceedings of the International Conference on Computing, Analytics, and Security Trends*, pp. 68–71, IEEE, Pune, India, 2016.
 - [19] S. Martinčić-Ipsić, E. Moćibob, and M. Perc, "Link prediction on twitter," *PLoS One*, vol. 12, no. 7, pp. 1–21, 2017.
 - [20] M. Jalili, Y. Orouskhani, M. Asgari, N. Alipourfard, and M. Perc, "Link prediction in multiplex online social networks," *Royal Society Open Science*, vol. 4, no. 2, pp. 1–11, 2017.
 - [21] Y. L. Shen, N. D. Thang, H. Y. Zhang, and T. T. My, "Interest matching information propagation in multiple online social networks," in *Proceedings of the 21th ACM International Conference on Information and Knowledge Management*, pp. 1824–1828, ACM, New York, NY, USA, 2012.
 - [22] Y. Liu and S. Xu, "Detecting rumors through modeling information propagation networks in a social media environment," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 46–62, 2016.
 - [23] X. G. He, M. Gao, M. Y. Kan, Y. Q. Liu, and K. Sugiyama, "Predicting the popularity of web 2.0 items based on user comments," in *Proceedings of the 37th International ACM and SIGIR Conference on Research and Development in Information Retrieval*, pp. 233–242, ACM, New York, NY, USA, 2014.
 - [24] I. N. Lymperopoulos, "Predicting the popularity growth of online content: model and algorithm," *Information Sciences*, vol. 369, pp. 585–613, 2016.
 - [25] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in Twitter," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pp. 355–358, Washington, DC, 2010.
 - [26] S. B. Kong, L. Feng, G. Z. Sun, and K. Luo, "Predicting lifespans of popular tweets in microblog," in *Proceedings of the 35th International ACM and SIGIR Conference on Research and Development in Information Retrieval*, pp. 1129–1130, ACM, New York, NY, USA, 2012.
 - [27] G. Haralabopoulos, I. Anagnostopoulos, and S. Zeadally, "Lifespan and propagation of information in on-line social networks: a case study based on Reddit," *Journal of Network and Computer Applications*, vol. 56, pp. 88–100, 2015.
 - [28] T. Britton, "Epidemic models on social networks with inference," *Statistica Neerlandica*, vol. 74, no. 2, pp. 222–241, 2020.
 - [29] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, Cambridge, UK, 2014.
 - [30] M. Hassanpour, S. A. Hoseinitabatabaei, P. Barnaghi, and R. Tafazoli, "Improving the accuracy of the video popularity prediction models through user grouping and video popularity classification," *ACM Transactions on the Web*, vol. 14, no. 1, pp. 1–28, 2020.
 - [31] H. V. Singh, "Predicting the popularity of online news using social features," in *Proceedings of the 2th International Conference on Green Computing and Internet of Things*, pp. 514–518, Bangalore, India, 2018.
 - [32] C. C. Hsu, L. W. Kang, C. Y. Lee, J. Y. Lee, Z. X. Zhang, and S. M. Wu, "Popularity prediction of social media based on multi-modal feature mining," in *IProceedings of the 27th ACM International Conference on Multimedia*, pp. 2687–2691, ACM, New York, NY, USA, 2019.
 - [33] J. Wan, X. Chen, Y. Du, and M. Jia, "Information propagation model based on hybrid social factors of opportunity, trust and motivation," *Neurocomputing*, vol. 333, pp. 169–184, 2019.
 - [34] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 88–97, IEEE, Sydney, Australia, 2011.
 - [35] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proceedings of the International Conference on*

- Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 67–75, Springer, Zagreb, Croatia, 2008.
- [36] B. Shulgin and L. Stone, “Pulse vaccination strategy in the SIR epidemic model,” *Bulletin of Mathematical Biology*, vol. 60, no. 6, pp. 1123–1148, 1998.
 - [37] M. Wang, A. L. M. Vilela, L. Tian, H. Xu, and R. Du, “A new time series prediction method based on complex network theory,” in *Proceedings of the IEEE International Conference on Big Data*, pp. 4170–4175, Boston, MA, USA, 2017.
 - [38] M. Khaksar Manshad, M. R. Meybodi, and A. Salajegheh, “A new irregular cellular learning automata-based evolutionary computation for time series link prediction in social networks,” *Applied Intelligence*, vol. 51, no. 1, pp. 71–84, 2021.
 - [39] Y. Wu, X. L. Chen, and Z. Y. Jiang, “Survey on predicting popularity of information in microblogs,” *Journal of Xihua University*, vol. 36, no. 1, pp. 1–6, 2017.
 - [40] C. Hu, Y. Hu, W. Xu, P. Shi, and S. Fu, “Understanding popularity evolution patterns of hot topics based on time series features,” *Web Technologies and Applications*, Springer, in *Proceedings of the Asia-Pacific Web Conference*, pp. 58–68, 2014.
 - [41] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, “Rise and fall patterns of information diffusion: model and implications,” in *Proceedings of the 18th ACM and SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 8710, pp. 6–14, ACM, New York, NY, USA, 2012.
 - [42] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, “TrendLearner: early prediction of popularity trends of user generated content,” *Information Sciences*, vol. 349–350, pp. 172–187, 2016.
 - [43] Y. Hu, C. Hu, S. Fu, P. Shi, and B. Ning, “Predicting the popularity of viral topics based on time series forecasting,” *Neurocomputing*, vol. 210, pp. 55–65, 2016.
 - [44] P. Qin and C. Cheng, “Prediction of seawall settlement based on a combined LS-ARIMA model,” *Mathematical Problems in Engineering*, vol. 2017, Article ID 7840569, 7 pages, 2017.
 - [45] Y. Zhong and G. Deng, “Geometric asian options pricing under the double heston stochastic volatility model with stochastic interest rate,” *Complexity*, vol. 2019, Article ID 4316272, 13 pages, 2019.
 - [46] H. Simon and W. John, “Kalman filtering and neural networks,” *Adaptive and Learning Systems for Signal Processing Communications and Control*, vol. 88, pp. 170–174, 2001.
 - [47] Q. Zhuang and L. Chen, “Dynamic prediction of financial distress based on kalman filtering,” *Discrete Dynamics in Nature and Society*, vol. 2014, Article ID 370280, 10 pages, 2014.
 - [48] J. A. Hartigan, *Bayes Theory*, Springer-Verlag, New York, NY, USA, 1983.
 - [49] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, “A stochastic quasi-Newton method for large-scale optimization,” *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1008–1031, 2016.
 - [50] Z. Mohammed, “Scalable algorithms for association mining,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 3, pp. 372–390, 2000.
 - [51] F. Par, D. G. Gasulla, A. Vilalta et al., “Fluid communities: a competitive, scalable and diverse community detection algorithm,” in *Proceedings of the 6th International Conference on Complex Networks and Their Applications*, pp. 229–240, Springer, Lyon, France, 2017.