IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2020, accepted March 21, 2020, date of publication March 26, 2020, date of current version April 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983583

A Prediction Method of Peak Time Popularity Based on Twitter Hashtags

HAI YU[1], YING HU[1], AND PENG SHI[2]
[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Peng Shi (pshi@ustb.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0803302.

ABSTRACT Understanding the peak time of popularity evolution can provide insights on recommendation systems and online advertising campaigns. Although popularity evolution has been largely studied, the problem of how to predict its peak time remains unexplored. Taking Twitter hashtags as case study, the goal of this study is to predict when popularity reaches the peak for Twitter hashtags, from the perspective of an online social network application, in the context of the Twitter social network. On the whole, this paper includes three research aspects. Firstly, this paper investigates how early popularity reaches its peaks. Then, it is found that popularity tends to peak in the early stage of its evolution. Secondly, this paper discusses when a peak time prediction should be triggered. Thirdly, this paper designs a multi-modal based deep learning method, where the state-of-art deep learning techniques, such as multi-modal embedding and attention mechanisms, are adopted. We find that in the early stage of popularity evolution, no matter which factor is used as the input, the prediction effect is poor. By contrast, the hashtag string factor has the weakest contribution to popularity prediction in the middle and late stages of popularity evolution. The overall performance of our proposed method is evaluated in terms of the minimum, quartiles, and maximum values of absolute errors. From the experimental results, the prediction method we designed is superior.

INDEX TERMS Popularity evolution, prediction, Twitter hashtags, attention mechanism, multi-modal deep learning.

I. INTRODUCTION

As massive amounts of online information are constantly being produced by social media sites, people are inundated with overloaded information. However, people's attention [1], [2] is a kind of scarce resource, which leads to the power-law distribution of information popularity [3], [4]. Most information received little attention, while some succeeded in achieving a huge bulk of popularity. For the popular information, a natural problem, how popularity evolves over time (popularity evolution), has been brought to the center [5], [6].

Popularity exhibits rich temporal variation [7] with spike [8], [9] reoccurring in the course of the evolution. Despite the variation of popularity evolution, popularity tends to accumulate in peaks [10]. For example, ninety percent of total popularity can sometimes be observed in a peak day or hour. It has also been revealed that the peak of popularity evolution has an exponential rise and power-law fall [8]. Peak fractions (the fraction of peak popularity compared with total popularity) can be used to group popularity evolution into endogenous or exogenous categories [10]. Therefore, the "peak" plays an important role in popularity evolution, since it is not only an inherent behavior of popularity evolution but also reveals people's intensive attention to online information. Given the importance of the "peak", an interesting question arises: Can we predict the peak time of popularity evolution?

Predicting the peak time of popularity evolution provides valuable insights for social media applications and services. For example, content providers can create a personalized timeline for users and online recommendation systems [11], [12] can promote articles or threads according to the peak time. For advertising, the peak time of popularity evolution is associated with the effectiveness of an advertising campaign.

At present, considerable works have been conducted on popularity evolution of online information. Various features, including early popularity [13]–[20], user comments [21], [22], network structures [23], and cascade information

The associate editor coordinating the review of this manuscript and approving it for publication was Mianxiong Dong.

[24]–[27], are extracted to predict popularity evolution [13]–[19], [21]–[23], [28], [29]. The popularity evolution was originally predicted based on empirical studies [30], [31], such as observing the strong correlation between early and late log-transformed popularity [13], or calculating time series similarities of popularity [15], [32] at different points in time. Recently, methods based on deep learning, especially multi-modal methods [18], [33]–[35] have been introduced into popularity evolution prediction. This is because social media sites, such as Facebook, Twitter, provide us with multi-source online data: text data, image data and network (graph) data, and multi-modal methods are able to make full use of those multi-source data. Text features extracted by Long Short-Term Memory (LSTM) [36]/Gated Recurrent Unit (GRU) [37] etc., image features extracted by Residual Neural Network (ResNet) or Xception etc., are combined in multi-modal modules [18], [35] to express the evolution of popularity. However, most existing work has focused on predicting popularity ''volume'' (e.g. predicting the value of popularity one day/hour later). In this paper, the problem is considered from the ''time'' angle: predicting the peak time of popularity evolution by taking Twitter hashtags as a case study. Furthermore, no work has taken network data (or graph data, users as nodes, following relationships between users as edges) as a modality in the multi-modal deep learning method, which motivates this paper to make use of topological network data, in addition to other multi-source data.

In this paper, we first conduct empirical studies to investigate how much time it usually takes popularity to reach its peak since popularity evolution begins. We find that popularity tends to reach its peak in the early stage of its evolution, which requires us to collect sufficient information from multiple data sources before predictions. Since the Twitter platform contains a variety of data sources, such as text data and network (graph) data, this paper would like to make full use of these data. Then, a multi-modal [33] deep learning solution is designed for predicting the peak time of popularity evolution. In this solution, LSTM and DeepWalk [38] are adopted for social information representation, hashtag string representation, and topological network representation. Next, the attention mechanism is applied. Finally, the attended and concatenated vector is sent into non-linear layers. The metric for this question is absolute error (the difference between the ground truth value and the predicted value). Experimental and comparative results show that the prediction solution we designed is superior.

This work is a follow-up of our previous study that provided a conventional machine learning method (Support Vector Regression, SVR) [39]. We here extend it by conducting some empirical studies and providing a multi-modal deep learning method with better performance.

II. RELATED WORK

A. POPULARITY EVOLUTION PREDICTION

In recent years, data-based prediction has become an important way for people to grasp the development and change trend of things, and relevant data prediction methods are emerging. Helbing *et al.* [40] discussed models and data on population disasters, crime, terrorism, war, and the spread of disease. They pointed out that complexity science was a way to better understand popularity prediction from the point of view of complex system. Wang *et al.* [41] studied mathematical models of disease transmission based on tools and concepts of statistical physics, and further with the help of new digital data sources, and proposed models that capture nonlinear interactions between behavior and disease dynamics to help to understand disease dynamics and inform prevention strategies. Similarly, there have been many efforts concentrating on popularity evolution prediction, ranging from popularity volume prediction [13]–[19], [21]–[23] to burst prediction [42], [43].

For popularity volume prediction, some researchers predicted the value of popularity in the near future, mainly based on two types of methods: the statistics-based method and the model-based method. In the statistics-based method [13], the correlation between early popularity and future popularity was learned through scatter plots and Pearson correlation analysis. In the model-based method [19], [44], stochastic process models were often utilized to characterize the process of how popularity is gained over time.

For burst prediction, some researchers predicted whether popularity would burst or not and when popularity would burst, mainly based on machine learning methods. Kong *et al.* [42] presented a binary classification task: will popularity burst in the near feature? Then, they found that the Support Vector Machine (SVM) model was the best tool in their task. Wang *et al.* [43] predicted the time of a burst. Due to the diverse time spans of popularity evolution, they formulated their problem as a classification problem to predict the time, when window burst will appear.

In contrast to most existing work, rather than predicting the future popularity volume, this paper considers the popularity evolution problem from the ''time'' angle: predicting the peak time of popularity evolution.

B. DEEP MULTI-MODAL LEARNING

With the flourishing of deep learning methodologies, some researchers have utilized deep multi-modal learning for popularity prediction [33]. Deep multi-modal learning involves three types of settings: multi-modal fusion, cross modality learning, and shared representation learning. (This paper mainly studies multi-modal fusion.)

Khosla *et al.* [34] analyzed the reasons why an image was popular. They investigated two features that might affect an image's popularity, namely the image content and social context. They first predicted the popularity of images using two features separately. They further fused two features and found that the fusion boosts prediction accuracy. Zhang *et al.* [35] fused text content and social context to make retweeting predictions. These features were converted to the representation of embedding with the attention mechanism firstly. The retweeting behavior was predicted through a fully connected

SoftMax function. Zhang *et al.* [18] predicted the popularity of social images by fusing visual features, textual features, and social features with VGGNet, LSTM, and the attention mechanism.

Most of these deep multi-modal papers made modalities for image features, text features and social information features. However, no work involves the topological network feature, which motivates this paper to consider topological network features. To predict peak time, we make multi-modal fusion by converting these factors into embedding with the DeepWalk algorithm, LSTM, and the attention mechanism. Next, the embedding is concatenated together and fed into fully connected layers. To the best of our knowledge, this is the first work to incorporate the topological network with the attention mechanism into multi-modal fusion for popularity prediction tasks.

III. PRELIMINARIES

A. DEFINITIONS

1) POPULARITY

By the popularity of a piece of online information, we refer to the amount of attention this information receives, such as, the number of views that a video receives, or the number of users discussing a hashtag. This paper takes the number of users discussing a hashtag as hashtag popularity.

2) POPULARITY EVOLUTION [45]

It is noted that most pieces of online information undergo both active and inactive periods [46]. We use the same method as that in [46] to distinguish between both periods: we consider a piece of information inactive if it fails to gain popularity for 24 hours. To simplify the problem, we shorten popularity evolution to the single active period, during which most popularity volumes accumulate. Given the observations of the popularity of a piece of online information i over its popularity evolution span L_i, $L_i \in N^+$, we define $y_i(t)$ as the popularity received by the piece of information i at time t, $t \in \{1, 2, 3, \ldots, L_i\}$. The data granularity is set to 1 hour. For example, $y_i(10)$ denotes the popularity received by i in the 10th hour. The popularity evolution of i is given by the time series $\{y_i(1), y_i(2), y_i(3), \ldots, y_i(L_i)\}$.

3) PEAK TIME

By peak time, we refer to the amount of time it takes popularity to reach the highest value once popularity evolution begins. The peak time of i is represented by T_i^p, $T_i^p \in \{1, 2, 3, \ldots, L_i\}$.

B. DATA SET

Our primary data comes from a portion of the 'tweet7' data set crawled by Yang and Leskovec over a period of 7 months from June to December 2009 [7]. (This data set complies with the terms of service for the Twitter website.) The data set comprises 65 million tweets. We identify 3.3 million hashtags in these tweets. From Figure 1 we can see that the popularity distribution of these 3.3 million hashtags follows a power-law shape. Most of the hashtags in our data set gain extremely small popularity whereas only a few hashtags gain large popularity. Since studying peaks requires hashtags with high peak popularity, we remove the hashtags with the low peak popularity volume.

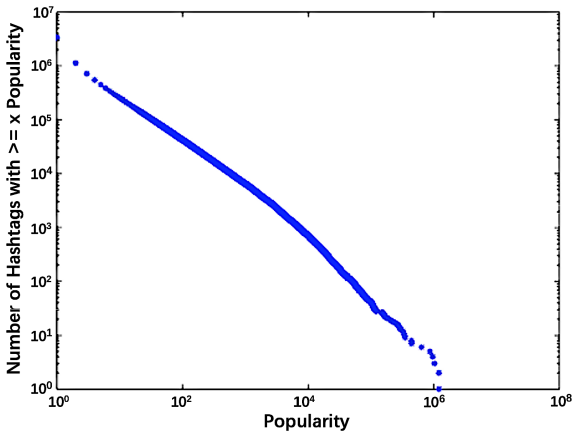

FIGURE 1. The popularity distribution of all hashtags in our dataset. The popularity distribution follows the power-law rule, which indicates that most of the hashtags in our data set gain very small popularity whereas only a few hashtags gain large popularity.

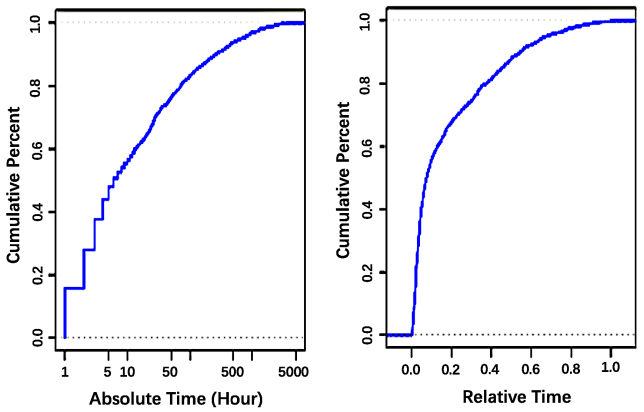

FIGURE 2. The empirical cumulative distributions of absolute and relative peak times.

C. CHARACTERISTICS OF PEAK TIME

In this section, we analyze how much time it usually takes popularity to reach its peak and in which stage of evolution (early, middle, or late stage) popularity usually peaks. The absolute peak time and relative peak time are adopted as metrics to answer this question. By absolute time, we refer to the number of hours it takes popularity to peak since popularity evolution begins. By relative time, we refer to the fraction of absolute peak time compared with evolution span.

Figure 2 shows the empirical cumulative distributions of absolute and relative peak times. It should be noted that for the left subplot, the horizontal axis is logarithmically rescaled because the large variances appear among absolute

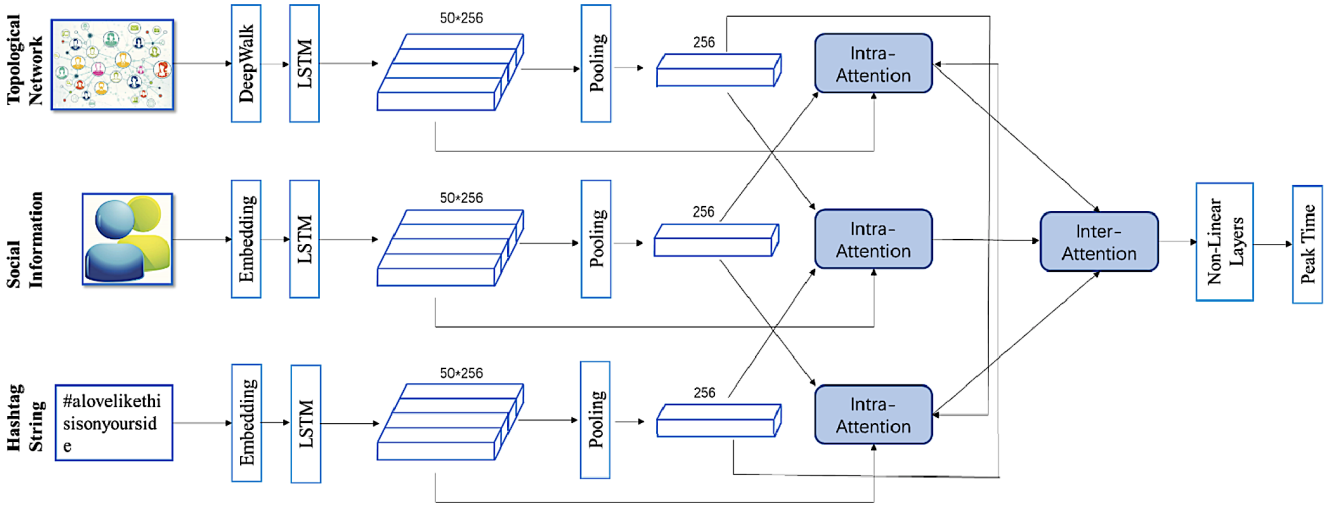

FIGURE 3. Architecture of the multi-modal based deep learning method. First, three types of data resources in Twitter social networks are incorporated, including the hashtag string information, social information, and topological network information. Second, various embedding algorithms, such as the DeepWalk algorithm and the LSTM model, are adopted to convert these three types of data resources to three 50*256-dimentional matrices (50 stands for the number of time steps of LSTM). Third, average pooling layers are applied to get vectors to make feature representations. Forth, the attention mechanism is adopted to learn intra-attention between any two of these three types of data resources. Fifth, inter-attention is learned among three intra-attentions. Finally, non-linear layer with one neural as the final output is used to predict the peak time of popularity evolution for a given hashtag.

peak times (they range from one hour to several thousand hours). For each observed point on a line in Figure 3, the y axis value shows the fraction of hashtags, for which their peak times do not exceed the corresponding x axis value.

The left subplot in Figure 2 indicates that about fifty-five percent, seventy-five percent, and ninety-five percent of hashtags experience their own peaks before the first 10, 50, and 500 hours, respectively. Some even experience their own peaks at the 1st hour, which indicates that popularity tends to peak quickly since evolution begins. The right subplot investigates the stage of evolution (early, middle, or late stages), in which popularity usually peaks. We can see that about sixty-eight percent, eighty-eight percent, and ninety-eight percent of all hashtags experience their own peaks in the first twenty percent, fifty percent, and eighty percent of their evolution spans, respectively. Therefore, popularity usually peaks in the early stage of its evolution.

Because popularity peaks in a short time, which may make the information we can obtain for prediction insufficient. Therefore, in order to collect sufficient prediction data, we make full use of multiple sources of data in the Twitter platform by multi-modal deep learning. In next section, we will present how to make peak time predictions for Twitter hashtags.

IV. PEAK TIME PREDICTION

In this section, we present how to use multi-modal deep learning to make peak time predictions for Twitter hashtags. Due to the limitations of our dataset, only three types of factors are incorporated. They are social information, hashtag strings, and topological network. We fuse these factors by making representations for them with deep learning techniques, such as LSTM and the attention mechanism. As shown in Figure 3, firstly, LSTM is used to make embedding for social information modality, hashtag string modality, and topological network modality, separately. Secondly, the proposed method utilizes intra-attention to learn attended embedding for these three modalities. Thirdly, inter-attention is adopted to learn different importance of different modalities. Finally, the learned multi-modal embedding is sent into non-linear layers for peak time predictions.

So, before diving into the proposed prediction method, we introduce mathematical notations. In this paper, we use bold letters to represent matrices and non-bold to represent vectors. For the hashtag i, given its peak time T_i^p, we have three representations as $\{\boldsymbol{S}^i, \boldsymbol{H}^i, \boldsymbol{N}^i\}$, where $\boldsymbol{S}^i, \boldsymbol{H}^i, and\ \boldsymbol{N}^i$ are representations of social information, hashtag strings, and topological network. Therefore, our target is to learn a function $f: \boldsymbol{S}^i, \boldsymbol{H}^i, \boldsymbol{N}^i \rightarrow T_i^p$. (What we need to say here is that this paper uses the terms: embedding and representation, interchangeably.)

A. MAKE EMBEDDINGS

1) HASHTAG STRINGS

A hashtag can be treated as a sequence of words. For example, the hashtag #alovelikethisisonyourside is interpreted as "a love like this is on your side". This paper converts each word to a word vector according to a pre-trained wiki text corpus. Therefore, the hashtag string can be represented as $\boldsymbol{H} = [w_1, \ldots, w_l]$, where l is the number of words of a hashtag string; the maximum length of l is 50 (as denoted in Figure 3), and w_t is a word vector.

Furthermore, LSTM (Long-short Term Memory) is adopted to encode hashtag string $\boldsymbol{H}.\{w_t\}_{t=1}^{l}$ is fed into LSTM.

At each time step, we have a hidden state h_t generated as follows.

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, w_t] + b_f) \quad (1)$$
$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, w_t] + b_i) \quad (2)$$
$$\tilde{C}_t = \tanh h(\boldsymbol{W}_c \cdot [h_{t-1}, w_t] + b_c) \quad (3)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$
$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, w_t] + b_o) \quad (5)$$
$$h_t = o_t * \tanh C_t \quad (6)$$

where f_t, i_t, and o_t are the forget gate, the input gate, and the output gate, respectively. σ is the sigmoid activation function. $\boldsymbol{W}$ and b are the parameters of LSTM. Hence, the final representation of hashtag strings can be denoted by $\bar{\boldsymbol{H}} = [h_1^H, \ldots, h_l^H]$.

2) THE TOPOLOGICAL NETWORK

The topological network can be represented as $\boldsymbol{N} = [N_1, \ldots, N_n]$, where n is the time point when a prediction is triggered. The time unit can be set to 1 hour. In this paper, N_t denotes the cumulative evolving network for a hashtag at time t, $t \in \{1, 2, \ldots, n\}$. The vertices of N_t are users who have tweeted on the hashtag in hours 0 through t. An edge between vertex u and vertex v is added if u and v have a follower-following relationship that comes from a data set collected by Kwak *et al.* [46]. The collection time of this data set is the same as that of the 'tweet7' data set. Then, in this paper, the DeepWalk [38] algorithm is utilized to make the embedding for N_t. Since N_t can have millions of vertices, we choose not to take every single vertex into account for the sake of computation speed in handling neural networks. In this paper, the top k significant vertices $(v_1, v_2, \ldots, v_k)$ in the network are chosen. N_t is given as follows.

$$N_t = [\phi(v_1), \phi(v_2), \ldots, \phi(v_k)] \quad (7)$$

where $\phi(v)$ is the embedding of the vertex v learned by the DeepWalk algorithm. ϕ is the mapping function. This mapping ϕ represents the latent social representation associated with the vertex v in N_t. According to the DeepWalk algorithm, $\phi(v)$ is calculated by the SkipGram [47] algorithm and random walks starting from v.

Furthermore, LSTM is adopted to encode the topological network. $\{N_t\}_{t=1}^n$ is fed into LSTM. At each time step, we have a hidden state h_t. Hence, the final representation of the topological network can be denoted by $\bar{\boldsymbol{N}} = [h_1^N, \ldots, h_n^N]$.

3) SOCIAL INFORMATION [48]

Social information involves the statistics of celebrities and fans. $\boldsymbol{S} = [u_1, \ldots, u_n]$, where n is the time point when a prediction is triggered, $u_t = \{num\_cel_t, tot\_fan_t, max\_fan_t, med\_fan_t, ave\_fan_t\}$. num_cel_t represents the number of celebrities involved until the time point t, $t \in \{1, 2, \ldots, n\}$. tot_fan_t, max_fan_t, med_fan_t, *and* ave_fan_t represent the overall sum, the maximum, the median, and the average of the number of fans of all users, respectively.

Furthermore, LSTM is adopted to encode social information. $\boldsymbol{S}.\{u_t\}_{t=1}^n$ is fed into LSTM. At each time step, we have a hidden state h_t. Hence, the final representation of social information can be denoted by $\bar{\boldsymbol{S}} = [h_1^S, \ldots, h_n^S]$.

B. INTRA-ATTENTION AND INTER-ATTENTION MECHANISMS

The intra-attention mechanism is presented to attend each modality for the hashtag string, the topological network, and social information embedding, respectively.

1) INTRA-ATTENTION FOR HASHTAG STRINGS

First, the topological network and social information embedding matrices are converted into vectors as follows.

$$\bar{n} = \frac{1}{n} \cdot \bar{\boldsymbol{N}}\vec{1} \quad (8)$$
$$\bar{s} = \frac{1}{n} \cdot \bar{\boldsymbol{S}}\vec{1} \quad (9)$$

where $\vec{1}$ is a vector with all elements being 1, and n is the number of time steps. The above equations can be interpreted as average pooling operations. After pooling operations, the representations of the topological network and social information are both vectors.

Second, we construct the hashtag string intra-attention score as follows.

$$\theta_{H,t} = \boldsymbol{W}_H(\tanh(\boldsymbol{W}_{HH} h_t^N) * \tanh(\boldsymbol{W}_{HN}\bar{n}) * \tanh(\boldsymbol{W}_{HS}\bar{s})) \quad (10)$$

where $\theta_{H,t}$ denotes the score of importance of a hidden state of hashtag string representation. tanh is adopted for squeezing values of embedding into the same range, which enhances non-linearity and avoids gradient exposure or vanishing. $\boldsymbol{W}_H$, $\boldsymbol{W}_{HH}$, $\boldsymbol{W}_{HN}$, and $\boldsymbol{W}_{HS}$ are parameters to be learned in attention layers. The above equation can be explained as the relevance of each hidden state of hashtag string representation to the topological network and social information representation jointly.

Finally, the attended hashtag string representation can be obtained as follows.

$$\alpha_H = \text{softmax}(\theta_H) \quad (11)$$
$$\check{h^H} = \sum_l \alpha_{H,t} \cdot h_t^H \quad (12)$$

where α_H is attention weights and $\check{h^H}$ is the attended hashtag string representation.

2) INTRA-ATTENTION FOR THE TOPOLOGICAL NETWORK

Likewise, the intra-attention for the topological network is given as follows.

$$\bar{h} = \frac{1}{l} \cdot \bar{\boldsymbol{H}}\vec{1} \quad (13)$$
$$\theta_{N,t} = \boldsymbol{W}_N(\tanh(\boldsymbol{W}_{NN} h_t^N) * \tanh(\boldsymbol{W}_{NH}\bar{h}) * \tanh(\boldsymbol{W}_{NS}\bar{s})) \quad (14)$$

TABLE 1. The minimum, quartiles, and maximum values of absolute errors. Looking at Q2 for Class 1 hashtags, the median absolute error is 1 if predictions triggered once popularity reaches 30. For Class 2 hashtags, it is 2. For all hashtags, it is 2, which indicates our solution can make peak time predictions with both good promptness and accuracy.

Stats	Class1					Class2					All				
	Min	Q1	Q2	Q3	Max	Min	Q1	Q2	Q3	Max	Min	Q1	Q2	Q3	Max
10	0	2	3	3	150	1	2	3	5	167	0	2	3	4	167
20	0	1	2	3	144	1	2	3	4	162	0	2	2	3	162
30	0	1	1	2	146	1	2	2	4	165	0	2	2	3	165
40	0	1	1	2	143	0	1	2	3	165	0	1	2	2	165

$$\alpha_N = \text{softmax}(\theta_N) \tag{15}$$
$$\check{h^N} = \sum_n \alpha_{N,t} \cdot h_t^N \tag{16}$$

3) INTRA-ATTENTION FOR SOCIAL INFORMATION

Likewise, the intra-attention for social information is given as follows.

$$\theta_{S,t} = \boldsymbol{W}_N(\tanh(\boldsymbol{W}_{SS}h_t^N) * \tanh(\boldsymbol{W}_{SH}\bar{s}) * \tanh(\boldsymbol{W}_{SN}\bar{n})) \tag{17}$$
$$\alpha_S = \text{softmax}(\theta_S) \tag{18}$$
$$\check{h^S} = \sum_n \alpha_{S,t} \cdot h_t^S \tag{19}$$

4) INTER-ATTENTION

The inter-attention mechanism is presented to capture different importance of these three modalities for different hashtags.

$$\theta_i = \tanh h(\boldsymbol{W}^i \check{h^l}), \quad i \in \{H, N, S\} \tag{20}$$
$$\alpha = \text{softmax}(\theta) \tag{21}$$

where θ denotes the vector of importance scores for these three modalities, and α denotes the vector of attention weights for these three modalities.

The attended multi-modal embedding r is computed as follows.

$$r = \alpha_H \check{h^H} + \alpha_N \check{h^N} + \alpha_S \check{h^S} \tag{22}$$

C. LEARNING FOR PEAK TIME PREDICTION

After obtaining the inter-attended multi-modal representation for hashtag strings, the topological network and social information, we adopt 2 fully connected layers to calculate peak time, which is given as follows.

$$\hat{T^p} = \boldsymbol{W}^2 \,\text{ReLU}\,(\boldsymbol{W}^1 r + b^1) + b^2 \tag{23}$$

where $\boldsymbol{W}^1$, $\boldsymbol{W}^2$, b^1, and b^2 are parameters to be learned in the fully connected layers. ReLU is the rectified linear unit, which enhances non-linearity for the model. $\hat{T^p}$ is the predicted peak time of popularity evolution of a given hashtag.

This paper formulates the peak time prediction task as a regression problem. Mean Square Error is minimized while training for the cost function, as shown below.

$$MSE = \frac{1}{n^T} \sum_{i=1}^{n^T} (T_i^p - \hat{T}_i^p)^2 \tag{24}$$

where n^T is the size of the training set; T_i^p is the ground truth peak time, and $\hat{T}_i^p$ is the predicted peak time for the hashtag i.

V. PREDICTION EVALUATION

In this section, we discuss how this prediction task is evaluated and present comparisons with baseline methods. What we need to say here is that this paper takes absolute error (the difference between the ground truth value and the predicted value) as the metric for Q3. All predictions are triggered before popularity reaches 40, as explained in [39].

The overall performance of our proposed method is evaluated in terms of the minimum, quartiles, and maximum values of absolute errors, as shown in Table 1. Where, according to whether popularity peaks in its first spike, we categorize peak times into two classes: the cases of popularity peaking in its first spike as Class 1, and the cases of popularity peaking in later spikes as Class 2. Meantime, Q1, Q2, and Q3 stand for the first quartile, the median, and the third quartile, respectively. To further illustrate this table, let's take the entry at the 4th row and the 5th column as an example. The value in this entry means that the third quartile of absolute errors is 3 for the prediction triggered when popularity reaches 20 for Class 1.

From Table 1, we can draw the following conclusions. (1) For the category All, median absolute error is 2 hours, if a prediction is triggered after popularity reaches 20, which indicates that our method has a good performance. (2) Predictions for Class 1 are more accurate than those for Class 2 because Class 1 hashtags experience earlier peaks of popularity evolution. Relatively, more sufficient data can be obtained for predictions for Class 1. (3) Overall errors decrease as predictions are triggered later, which is consistent with the intuition that the later we predict, the more accurate the prediction is. (4) The maximum absolute errors are large (greater than 150), which results from the hashtags experiencing peaks later than the 500th hour.

In addition, we find that at different stages of popularity evolution, different factors have different effects on prediction accuracy. It can be found by the overall comparison, in the early stage of popularity evolution, no matter which factor is used as the input, the prediction effect is poor, indicating that they have little influence on the prediction accuracy. When we trigger predictions later, the prediction improvement is obviously affected by time-variant factors which are the topological network factor and the social information factor.

Of course, we also find that in this improvement, the topological network factor is not as effective as we may think in predicting activity periods, even if the prediction is triggered late in the popularity evolution and there is enough historical data can be used. The social information factor is more important than other factors. The reason is that the more celebrities, or the more fans they have, the longer the active period of evolution will be and the better for us to predict.

By contrast, the hashtag string factor has the weakest contribution to popularity prediction in the middle and late stages of popularity evolution. This is because the hashtag string is a static factor, it does not change with time, and its embedding vector is constant no matter when the prediction is made.

Although the hashtag string embedding is time-invariant and does not contribute to this prediction improvement, we still incorporate the hashtag string embedding, since the hashtag is one of the most important resources that we can make use of in the Twitter social network, and previous studies [49], [50] have shown that the information of hashtags is effective for popularity prediction.

At present, there are many methods of popularity prediction, including three main categories. These are early popularity prediction [13], [51], [52], influence factor prediction [22], [53] and cascade propagation prediction [26], [54], [55]. To validate the effectiveness of our prediction method, we need to compare it with baseline methods. This paper selects several typical popularity prediction methods and compares our method with them in terms of absolute error. (Predictions are triggered when popularity reaches 20.)

A. NAM (NO ATTENTION MECHANISM)

For showing the effectiveness of the attention mechanism, this section compares the methods with and without the attention mechanism. For the method without the attention mechanism, we concatenate representation vectors of three modalities right after pooling layers. Next, concatenated vectors are directly fed into non-linear layers.

B. SVR

As we did in previous work [39], this paper feeds SVR handcrafting features, such as network topological features (the average node degree, the maximum node degree, the global clustering coe?cient, etc.) and hashtag string features (the number of words and hashtag string lengths).

C. SpikeM

Most of the existing models for popularity evolution predictions are incapable of solving this task, because they are designed for predicting the future popularity volume. To compare our solution with existing work, this section chooses the SpikeM model which can solve this task but is not specialized for it [42]. This paper trains the SpikeM model by using historical popularity data starting from the beginning of popularity evolution up to the prediction trigger time. Then peak times can be inferred from predicted popularity data.

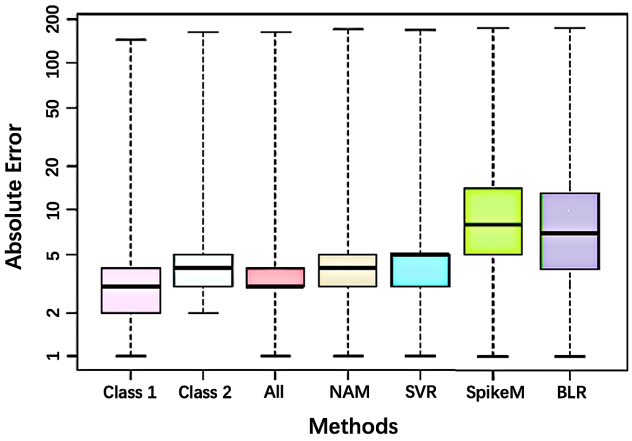

FIGURE 4. Absolute error comparison. Looking at the median absolute error, our solution's is lower than those of alternative methods, which shows the effectiveness of high-level features that are learned by multi-modal deep learning.

D. BLR (BAYESIAN LINEAR REGRESSION) [56]

It is another machine learning model corresponding to SVR. In this paper, the historical popularity data from the beginning of the popularity evolution to the trigger prediction moment are used as the training data of BLR, and then the peak moment is inferred from the predicted popularity evolution.

The absolute errors for predictions adopting our methods and alternative methods are shown in Figure 4. A box-and-whisker plot shows the minimum, quartiles, and maximum absolute errors. The bottom and the top of the box are the first and third quartile absolute errors, respectively, and the band inside the box is the median absolute error. The upper and lower whiskers are the maximum and minimum absolute errors, respectively.

The first three boxes show predictions using our method for Class 1, Class 2, and All hashtags. For better visualization, each prediction error is increased by one hour, and then presented on the logarithmically rescaled vertical axis.

We can draw the conclusions as follows. (1) Comparing the third and fourth boxes, the median absolute error for predictions with attention mechanism which is 3.4 is lower than the median absolute error for predictions without attention mechanism which is 4.5. This shows the effectiveness of our intra-attention and inter-attention mechanism. (2) Comparing the third, fifth, sixth and seventh boxes, we can see that the median absolute errors are respectively 3.4, 4.6, 8.0, 7.6. This shows the method based on deep learning is better than the method based on conventional machine learning (SVR, BLR) and also better than the method based on the suspectible-infected method (SpikeM). It also shows the effectiveness of high-level features that are learned by deep learning. In fact, sometimes the performance difference between machine learning and deep learning algorithms is not big, but the characteristics used play a decisive role. The later the prediction is triggered, the smaller the difference between the four methods. (3) Comparing the first three boxes, the median

absolute errors are respectively 3.0, 4.8, 3.4. This shows that the predictions for Class 1 hashtags are more accurate than the predictions for Class 2 hashtags. Predictions for all hashtags are compromised between the previous two. This is because Class 1 hashtags experience earlier peaks of popularity evolution. Relatively more sufficient data can be obtained for predictions for Class 1.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, the peak time of popularity evolution for Twitter hashtags is investigated. First, we analyze how early popularity reaches its peak, and find that popularity tends to peak quickly in the early stage of its evolution. Next, we present a multi-modal based deep learning method of making peak time predictions. Several deep learning techniques, such as the DeepWalk algorithm and LSTM, are adopted for multi-modal embedding of hashtag strings, social information, and the topological network. Then, the intra-attention and inter-attention mechanisms are learned for multi-modal embedding. Finally, experimental results show that our prediction method outperforms baseline methods.

In future work, we will investigate the peak time prediction method on other kinds of datasets, like YouTube video and Instagram picture datasets, and then incorporate more high-level features like visual features.

ACKNOWLEDGMENT

The authors give special thanks to Jure Leskovec and Sebastien Ardon for providing us with the Twitter dataset.

REFERENCES

[1] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 45, pp. 17599–17601, Nov. 2007.
[2] F. Wu and B. A. Huberman, "Popularity, novelty and attention," in *Proc. 9th ACM Conf. Electron. Commerce*, 2008, pp. 240–245.
[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, You Tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2007, pp. 1–14.
[4] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, no. 5, pp. 323–351, Sep. 2005.
[5] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, "On the dynamics of social media popularity: A YouTube case study," *ACM Trans. Internet Technol.*, vol. 14, no. 4, pp. 1–23, Dec. 2014.
[6] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: Characterizing popularity growth of YouTube videos," in *Proc. 4th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2011, pp. 745–754.
[7] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. 4th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2011, pp. 177–186.
[8] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: Model and implications," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 6–14.
[9] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, "Do cascades recur?" in *Proc. 25th Int. Conf. World Wide Web WWW*, 2016, pp. 671–681.
[10] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 41, pp. 15649–15653, Oct. 2008.
[11] M. Eirinaki, J. Gao, I. Varlamis, and K. Tserpes, "Recommender systems for large-scale social networks: A review of challenges and solutions," *Future Gener. Comput. Syst.*, vol. 78, pp. 413–418, Jan. 2018.
[12] X. Ma, J. Ma, H. Li, Q. Jiang, and S. Gao, "ARMOR: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks," *Future Gener. Comput. Syst.*, vol. 79, pp. 82–94, Feb. 2018.
[13] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.
[14] B. Chang, H. Zhu, Y. Ge, E. Chen, H. Xiong, and C. Tan, "Predicting the popularity of online serials with autoregressive models," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2014, pp. 1339–1348.
[15] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of YouTube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2013, pp. 365–374.
[16] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Perform. Eval.*, vol. 68, no. 11, pp. 1037–1055, Nov. 2011.
[17] S. N. Firdaus, C. Ding, and A. Sadeghian, "Topic specific emotion detection for retweet prediction," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2071–2083, Aug. 2019.
[18] W. Zhang, W. Wang, J. Wang, and H. Zha, "User-guided hierarchical attention network for multi-modal social image popularity prediction," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1277–1286.
[19] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, "Expecting to be HIP: Hawkes intensity processes for social media popularity," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 735–744.
[20] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1882–1895, Sep. 2016.
[21] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proc. Int. Conf. Web Intell., Mining Semantics (WIMS)*, no. 67, 2011, pp. 1–8.
[22] X. He, M. Gao, M.-Y. Kan, Y. Liu, and K. Sugiyama, "Predicting the popularity of Web 2.0 items based on user comments," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 233–242.
[23] M. X. Hoang, X.-H. Dang, X. Wu, Z. Yan, and A. K. Singh, "GPOP: Scalable group-level popularity prediction for online content in social networks," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 725–733.
[24] C. Li, J. Ma, X. Guo, and Q. Mei, "DeepCas: An end-to-end predictor of information cascades," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 577–586.
[25] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. 23rd Int. Conf. World Wide Web (WWW)*, 2014, pp. 925–936.
[26] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 2335–2338.
[27] M. Jalili and M. Perc, "Information cascades in complex networks," *J. Complex Netw.*, vol. 5, no. 5, pp. 665–693, Jul. 2017.
[28] S. Asur, B. A. Huberman, and G. Szabo, "Trends in social media: Persistence and decay," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 434–437.
[29] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, "TrendLearner: Early prediction of popularity trends of user generated content," *Inf. Sci.*, vols. 349–350, pp. 172–187, Jul. 2016.
[30] V. R. B. G. Caldiera and H. D. Rombach, "Goal question metric paradigm," *Encyclopedia Softw. Eng.*, vol. 1, pp. 528–532, 1994.
[31] M. Perc, "The Matthew effect in empirical data," *J. Roy. Soc. Interface*, vol. 11, no. 98, Sep. 2014, Art. no. 20140378.
[32] S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on Twitter," *PLoS ONE*, vol. 12, no. 7, 2017, Art. no. e0181079.
[33] N. Jiquan, K. Aditya, and K. Mingyu, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.
[34] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proc. 23rd Int. Conf. World Wide Web (WWW)*, 2014, pp. 867–876.
[35] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, "Retweet prediction with attention-based deep neural network," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2016, pp. 75–84.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *J. Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] Y. Wang, W. Liao, and Y. Chang, "Gated recurrent unit network-based short-term photovoltaic forecasting," *Energies*, vol. 11, no. 8, p. 2163, 2018.
[38] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 701–710.
[39] Y. Hu, C. Hu, S. Fu, M. Fang, and W. Xu, "Predicting key events in the popularity evolution of online information," *PLoS ONE*, vol. 12, no. 1, 2017, Art. no. e0168749.
[40] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, and M. Perc, "Saving human lives: What complexity science and information systems can contribute," *J. Stat. Phys.*, vol. 158, no. 3, pp. 735–781, Feb. 2015.
[41] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, "Statistical physics of vaccination," *Phys. Rep.*, vol. 664, pp. 1–113, Dec. 2016.
[42] S. Kong, Q. Mei, L. Feng, F. Ye, and Z. Zhao, "Predicting bursts and popularity of hashtags in real-time," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 927–930.
[43] S. Wang, Z. Yan, and X. Hu, "Burst time prediction in cascades," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 325–331.
[44] Q. Zhao, M. A. Erdogdu, and H. Y. He, "Seismic: A self-exciting point process model for predicting tweet popularity," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1513–1522.
[45] Y. Hu, C. Hu, S. Fu, and J. Huang, "Survey on popularity evolution analysis and prediction," *J. Electron. Inf. Technol.*, vol. 39, no. 4, pp. 805–816, 2017.
[46] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 591–600.
[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[48] K. Lerman, "Social information processing in news aggregation," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 16–28, Nov. 2007.
[49] O. Tsur and A. Rappoport, "What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 643–652.
[50] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in Twitter," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 7, pp. 1399–1410, Jul. 2013.
[51] W. O. Kermack and A. G. Mckendrick, "A contribution to the mathematical theory of epidemics," *Proc. Roy. Soc. London. A, Containing Papers Math. Phys. Character*, vol. 115, no. 772, pp. 700–721, 1927.
[52] W. O. Kermack and A. G. Mckendrick, "Contributions to the mathematical theory of epidemics. II.—The problem of endemicity," *Proc. Roy. Soc. London. A, Containing Papers Math. Phys. Character*, vol. 138, no. 834, pp. 55–83, 1932.
[53] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *Proc. 6th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, vol. 12, May 2012, pp. 26–33.
[54] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu, "On popularity prediction of videos shared in online social networks," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. CIKM*, 2013, pp. 169–178.
[55] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng, "Popularity prediction in microblogging network: A case study on Sina Weibo," in *Proc. 22nd Int. Conf. World Wide Web (WWW Companion)*, 2013, pp. 177–178.
[56] W. Gero and A. Thomas, "Bayesian linear regression," in *Monograph of Linear Regression Analysis*. Singapore: World Scientific, 2009, pp. 297–316.

HAI YU is currently pursuing the Ph.D. degree with the University of Science and Technology Beijing. His research interests are mainly in social networks, software engineering, and big data application in police field. He is also involved in the National Grand Fundamental Research 973 Program of China: The Study of Online Social Network.

YING HU received the Ph.D. degree from the University of Science and Technology Beijing. Her areas of interest include data mining and machine learning. She had been involved in the National Grand Fundamental Research 973 Program of China: The Study of Online Social Network.

PENG SHI received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2007. He is currently a Professor of computer science with the University of Science and Technology Beijing. He has published a book on community detection of network in Chinese, in 2008. He has also published more than 30 scientific articles. His research interests are mainly in social networks, knowledge engineering, and big data application in materials science.