Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features

Yingdan Shang[1], Bin Zhou[2]*, Xiang Zeng[1], Ye Wang[1], Han Yu[1] and Zhong Zhang[1]

[1]College of Computer Science and Technology, National University of Defence Technology, Changsha, China, [2]Key Lab of Software Engineering for Complex Systems, School of Computer, National University of Defence Technology, Changsha, China

Predicting the popularity of online content on social network can bring considerable economic benefits to companies and marketers, and it has wide application in viral marketing, content recommendation, early warning of social unrest, etc. The diffusion process of online contents is often a complex combination of both social influence and homophily; however, existing works either only consider the social influence or homophily of early infected users and fail to model the joint effect of social influence and homophily when predicting future popularity. In this study, we aim to develop a framework to unify the social influence and homophily in popularity prediction. We use an unsupervised graph neural network framework to model nondirectional social homophily and integrate the attention mechanism with the graph neural network framework to learn the directional and heterogeneous social relationship for generating social influence representation. On the other hand, existing research studies often overlook the social group characteristics of early infected users, and we try to divide users into different social groups based on user interest and learn the social group representation from clusters. We integrate the social influence, homophily, and social group representation of early infected users to make popularity predictions. Experiments on real datasets show that the proposed method significantly improves the prediction accuracy compared with the latest methods, which confirms the importance of joint model social influence and homophily and shows that social group characteristic is an important predictor in the popularity prediction task.

Keywords: popularity prediction, deep learning, social network analysis, social group, social influence, information diffusion, social homophily

1 INTRODUCTION

The rapid development of social network platforms, such as Twitter, Sina Weibo, etc., promotes the wide spread of online content among people. Once published, online content (such as a tweet, a microblog, or a video) spreads among users through reposting, citing, or commenting mechanisms on social network platforms. The popularity of online content is usually measured by the number of users participating in its diffusion process [1,2], and users who participate in the propagation of information are also referred to as activated or infected users, which are terms from epidemiology. Although most online content can only attract a small number of users but also some online content manage to attract many users and become very popular. Popular contents generally have a greater impact on society, for example, some widespread fake news may cause a decline in government credibility and social instability. Predicting the popularity of online contents in social networks has

always been a hot research topic and has been widely applied in many real-world applications, such as campaign strategies [3], online content recommendation [4], and viral marketing [5].

However, predicting the future popularity of online content is a challenging task because the information diffusion process is affected by many complex factors and fraught with uncertainty. Existing studies have shown that popularity prediction can be made with higher accuracy with the observation of the early spread process compared with ex-ante approaches [6,7] as ex-ante approaches make predictions before the publication of online content. Therefore, recent popularity prediction approaches generally try to extract predictive features from the early diffusion episodes and model the nonlinear relationships between early features and future popularity of online content. Previous studies [6,8,9] have demonstrated that features in the early diffusion episodes including content features, user features, diffusion patterns of popularity, structural features, etc., are closely related to the ultimate fate of online items. Recently, deep learning–based approaches learn data-driven diffusion models from the early diffusion episodes, avoiding heavy feature extraction work and strong propagation hypotheses, while keeping satisfactory prediction results.

As a pioneer of deep learning–based approaches to popularity prediction, DeepCas [10] is an end-to-end deep learning method which uses the random walk method to sample user sequence from cascade graphs. A cascade graph is formed by the users and their interaction relationship during the diffusion process. CasCN [11] learns the representation of each cascade subgraph through a graph convolution network (GCN) and then uses the structure representation of the subgraph as the input to long short-term memory (LSTM) to capture the dynamic evolution diffusion process. VaCas [12] uses GraphWave, an unsupervised spectral graph wavelet method, to learn the node representation in the cascade graph. CasFlow [13] is an extension work of VaCas, different from VaCas which only models cascade graphs. CasFlow constructs a global retweet graph based on the retweet relations among users and uses an extensible graph representation learning method—sparse matrix factorization [14] to learn user preferences from the global graph.

However, prior works consider either the social influence or homophily of early infected users and neglect to model the joint effect of social influence and homophily in popularity prediction. The dynamic diffusion process is often a complex combination of social influence and homophily [15–17]. It is often impossible to distinguish whether the diffusion is caused by homophily or social influence. Homophily suggests that similar individuals are more likely to interact with each other, which can induce correlated diffusion actions without direct causal influence. Intuitively, the social influence of the propagators determines which users the information is visible to, but the preference of the user largely determines whether the user will participate in the dissemination of information, such as forwarding or commenting, after seeing the information. Therefore, our key objective is to develop a data-driven framework to model the joint effect of social influence and homophily in the early diffusion process for popularity prediction.

The social influence of early adopters, that is, users who view or forward online content in the early stage of content diffusion is an important factor in determining the future scale of online content [8,18,19]. Intuitively, if some early adopters have a large number of fans, the online content will be visible to more people, thus leading to higher popularity of the online content. Most existing approaches model the social influence of users based on the structural information of the underlying follow-up network which is also known as a social network. In many cases, it is difficult for us to obtain the social network because of the privacy protection policies of social network platforms. In addition, the spread of information does not strictly flow along with the underlying social network topology. Since social behaviors such as retweeting between users are the direct manifestation of social influences and homophily, we construct the global retweet network based on the retweet relationships among users.

We use two different graph neural network frameworks to separately model the homophily and social influence of users since homophily between users is nondirectional, whereas social influences of users are directional. We use variational graph auto-encoders (VGAE) [20] to learn the nondirectional social homophily representation of users from the constructed global retweet network. Moreover, social ties can be strong or weak, and we use an attention-based graph neural network framework to learn the heterogeneous social strength in the network when generating social influence representation of users.

On the other hand, existing popularity prediction approaches often neglect the social group characteristics of early adopters. Existing studies [21–23] have shown that the diversity of communities involved in the early diffusion of information is directly proportional to the final popularity of the information. That is, the more communities the information affected, the more viral it will be in the future, where communities refer to subsets of nodes with dense connections in the social network. Users in social networks spontaneously form different social groups or communities according to the topics they are interested in. Therefore, we attempt to tackle this problem by clustering users into different social groups based on user interests and model the social group characteristics of early adopters for popularity prediction.

The main contributions of this study include the following points:

- We propose a novel framework to model the joint effect of social influence and homophily in the early diffusion process for popularity prediction when the social network is unknown.
- We utilize variational graph auto-encoders (VGAE) to learn the nondirectional social homophily representation and an attention-based graph neural network framework to learn the directional social influence representation.
- We divide users into different social groups based on user interest representation and learn the social group characteristics of early adopters for popularity prediction.
- Experimental results on two real datasets show that the proposed model has better prediction accuracy than the

other latest methods and demonstrates the effectiveness of our proposed method.

2 BACKGROUND AND RELATED WORKS

Many research studies have focused on predicting the popularity of online information and are broadly divided into three categories: feature-based approaches, generative approaches, and deep learning–based approaches. Feature-based approaches identify and extract hand-crafted predictive features, including temporal features [24,25], structural features [22,26], and content features [27,28], and use machine learning algorithms such as linear regression and support vector machines to make predictions. This type of approach generally requires heavy feature engineering and the performance is strongly determined by the effectiveness of extracted features. Generative approaches mainly model the diffusion process of online content by probabilistic statistical generative approaches, for example, epidemic models [29] and point processes [30–32]. The underlying rationale behind this type of approach is that future popularity can be predicted by the temporal information in the early diffusion episode. The performance is normally limited by its strong assumptions of underlying diffusion mechanisms.

Inspired by the successful application of deep learning methods in natural language processing, image processing, and other fields, more researchers use deep learning models to automatically make predictions based on the representation of information diffusion features [33–35]. Existing deep learning–based approaches can avoid time-consuming feature engineering and have achieved significant improvement in prediction accuracy. We introduce the related approaches according to their underlying rationale as follows.

2.1 Social Influence
Traditional feature–based methods generally use the number of fans [18], PageRank [36], historical amount of retweeting [37], etc., to measure the social influence of users. The study [8] tracks the spread of a large number of tweets and finds that tweets of high popularity are usually generated by users with large influence and followers. The study [18] comes to the same conclusion that the final popularity of online information is closely related to the social influence of users. In addition, the self-excited Hawkes process [38] is one of the generative approaches which consider the effect of endogenous social influence–triggered forwarding on the microblogging platform. However, a major limitation on the parameterization of diffusion processes is their limited ability to express arbitrarily distributed data and are often oversimplified in practical applications.

As a classical deep learning-based approach, DeepHawkes [39] uses a gated recurrent unit (GRU) to obtain the user representation from the diffusion path and takes the time decay effect into account when generating cascade embedding. FOREST [40] proposes a multi-scale diffusion prediction model using reinforcement learning which can simultaneously predict the macroscopic popularity and the microscopic next infected users. Coupled-GNN [41] uses two mutually coupled GNNs to

model early adopters, where one GNN models the mutual influence of users and the other models the activation state of users, and the results are coupled through the gate mechanism. However, the aforementioned two methods require the underlying social graph. In many practical application scenarios, reliable social networks cannot be accessed because of the privacy protection policy of platforms.

Liu [42] tries to make predictions when the social network structure and the interaction relationship are both unknown, and the study helps in constructing the interaction graph and uses the infection time interval as the weight of edges. This artificial assumption does not conform to the facts and may have a certain negative impact on the prediction accuracy. CasFlow [13] uses the GraphWave method to model the social influence of users in the cascade graph; however, this model does not consider the direction of user influence because the Laplacian matrix cannot be applied to directed graphs.

2.2 Homophily and Social Group
The structural characteristic of social networks, such as the community structure, is a reflection of homophily. It is found that the community structure has a remarkable impact on the dynamics of information diffusion [21,22,43]. The study [44] finds that there is an optimal community structure that can significantly promote information propagation. Weng [21] analyzed Twitter datasets and showed that popular online contents tend to spread like epidemics, and viral memes permeate more communities than less viral memes. Furthermore, Weng [22] confirmed that we can predict future popularity based on the community structure information in the early spread process. Liu [23] adopted the widely used community detection method such as Infomap to compute community labels for users using global social networks and then used an RNN-based neural network to model the user community and temporal information; experiments on real-world datasets demonstrate that community information of early infected users can enhance the precision of predicting the next activated user.

On the other hand, homophily reflects the similarity of user preferences and behaviors. User preference in the social network can be expressed by the semantic information of the text they participate in. Research [45] regards the user's influence as an external influencing factor and the user's interest level in the topic as an internal influencing factor. TD-IDF is used to get the textual representation of the information, and LDA is used to obtain the topic representation of the information; the model calculates the similarity between user interest and information content and uses machine learning models such as SVM and Random Forest to predict the users who may be active in the future. Zhang [46] uses a convolutional neural network (CNN) to learn the semantic representation of tweets, learns the representation of user interest from the user's historical tweets, and then predicts the users who may be infected based on the similarity between the user's interest and the tweet. T-DeepHawkes [47] concatenates the cascade embedding which is generated by DeepHawkes and the text embedding of online content generated using LDA to make a popularity prediction. The model focuses on the semantic information of online content rather than user homophily or
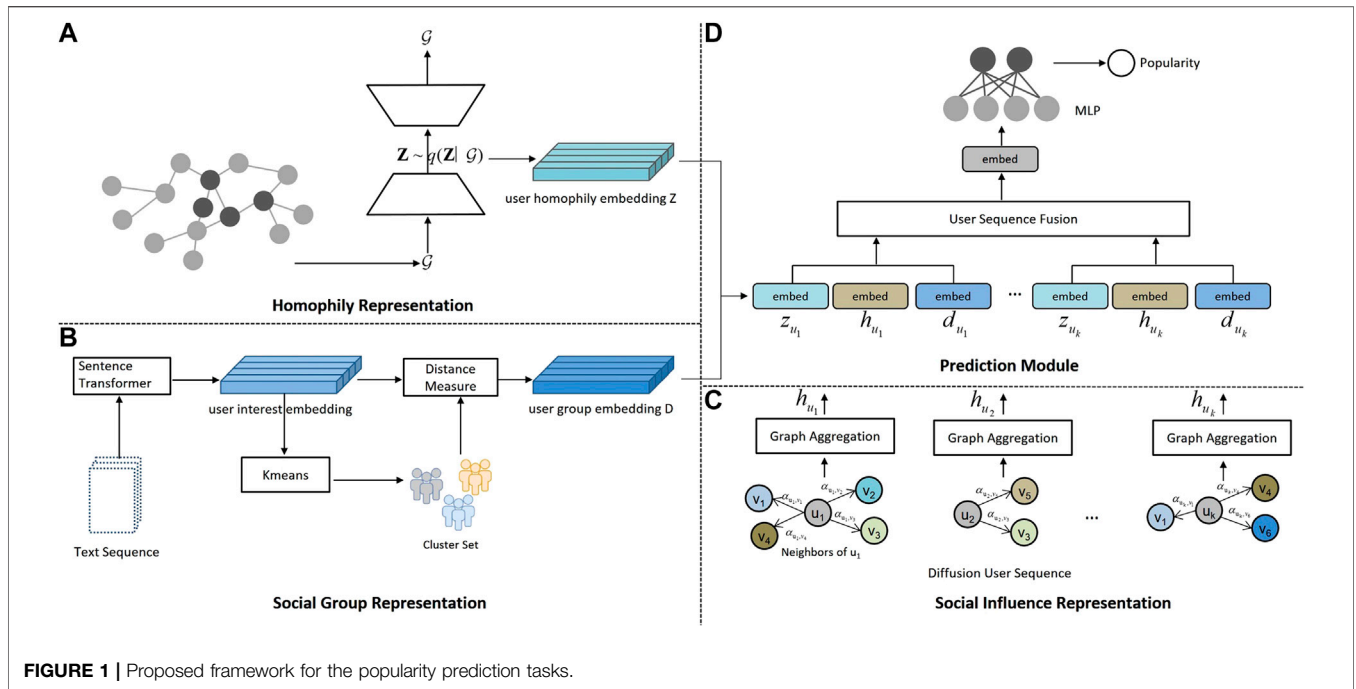
FIGURE 1 | Proposed framework for the popularity prediction tasks.

social group characteristic. TAN [48] incorporates topic information of online content and user dependencies into user representations *via* attention mechanisms to predict the next infected users.

However, the aforementioned research studies which model the homophily and social group features of early adopted users generally focus on the micro-level prediction which predicts the next infected users and neglect its importance in popularity prediction tasks.

3 METHODS

3.1 Problem Definition

The popularity of online information is usually defined as the attention level of social network users to the information. Consistent with many previous works [11,39], we formulate the task of popularity prediction as a regression task, which aims at predicting the number of users who will participate in the spread of online information in the future.

Suppose the given social network is composed of N users $\mathcal{U} = (u_1, u_2, \ldots, u_N)$, we use $D(T)$ to represent the sequence of users who have been infected before observation time T: $D(T) = \{u \in \mathcal{U} | t_u < T\}$, t_u is the infected timestamps of user u. Our task is to predict the number of all eventually infected users | $D(t_\infty)|$. Given a set of m observed diffusion episodes of online information before the observation time $\mathcal{D}(T) = \{D_i(T) | i = 1, 2, \ldots, m\}$, we extract the following information from the episodes:

1. The global retweet network $\mathcal{G}$, which represents the retweet relationship among social network users and is represented by

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices which represents social network users in $\mathcal{U}$ and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges which denotes the interaction relationship (e.g., retweet, comment) between users. For example, if user v retweets a certain online information published by user u, there is a directed edge (u, v) between nodes u and v.

2. Historical text sequences of users $\mathcal{M}$. A user may participate in multiple diffusion episodes of online information. We use $m_u = \{s_1, s_2, \ldots s_k\}$ to denote the historical text sequence of user u, where s_i represents the text content of online information. We use $\mathcal{M} = \{m_1, m_2, \ldots m_N\}$ to represent all the users and their corresponding historical text sequence.

We aim to learn a popularity prediction model by using the diffusion data before the observation time to predict the final incremental popularity P of online content: $f: \{\mathcal{D}, \mathcal{G}, \mathcal{M}\} \rightarrow \Delta P(t_f)$, where $P(t_f) = |D(t_\infty)| - |D(T)|$ represents the incremental popularity of online content.

3.2 Models

This section describes our work to model user influence and user group characteristics for predicting the eventual popularity of online content. Our model can be divided into four main parts as **Figure 1** shows: 1) the homophily representation module, which uses the variational graph auto-encoder (VGAE) to learn the user's homophily representation from the constructed global retweet network; 2) the social group representation module, which uses the natural language processing model to obtain the user interest representation according to the historical text sequence of users. Based on the user interest representation, social network users are divided into different social groups, and the distance between users and group centers is used as the

representation of social group features. 3) The social influence representation module, which uses the graph neural network and attention mechanism to learn the user's social influence representation from the directed global retweet network. 4) The prediction module, which integrates the social influence, homophily, and social group representation of early users who participate in the early period of information diffusion episodes for predicting the future popularity of online content.

3.2.1 Homophily Representation

Homophily describes the phenomenon that individuals with similar interests tend to associate with one another. The user's own preference largely determines whether the user will participate in the dissemination of information. The structural features of social networks such as the community structure reflect the homophily of users, moreover, homophily is directionless. Therefore, we use the variational graph auto-encoder (VGAE) to extract the nondirectional structural features which reflect the user's homophily from the global retweet network.

In the VGAE framework, we use a two-layer graph convolutional network (GCN) as the encoder and an inner product as the decoder. The GCN neural network extracts features for each node by looking at its neighbor nodes and summarizes them into low-dimension latent vectors. The inner product decoder is used to reconstruct the graph adjacency matrix. For our global retweet network $\mathcal{G}$, we define its adjacency matrix as A and the node feature matrix as X. Initially, each node in the graph is represented as a one-hot vector, $q \in \mathbb{R}^N$, where N is the total number of nodes in the graph. All nodes are then converted to a low-dimensional dense representation by sharing an embedding matrix E: $x = qE, E \in \mathbb{R}^{N \times D}$, where D is an adjustable dimension.

First, we use the adjacency matrix and node feature matrix as the input and generate the matrix of mean vectors and standard deviation vectors through a two-layer graph convolutional network:

$$GCN\left(\mathbf{X}, \mathbf{A}\right) = \tilde{\mathbf{A}} ReLU\left(\tilde{\mathbf{A}} \mathbf{X} \mathbf{W}_0\right) \mathbf{W}_1 \qquad (1)$$

$$\boldsymbol{\mu} = GCN_{\boldsymbol{\mu}}\left(\mathbf{X}, \mathbf{A}\right), \boldsymbol{\sigma} = GCN_{\boldsymbol{\sigma}}\left(\mathbf{X}, \mathbf{A}\right) \qquad (2)$$

$\mathbf{W}_0$ and $\mathbf{W}_1$ are trainable weight matrices of the GCN, $ReLU()$ is the non-linear function, and $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix. We sample the latent vectors of each node from normal distribution: $q(z_i \mid \mathbf{X}, \mathbf{A}) = \mathcal{N}(z_i \mid \boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2))$, and get the representations for all nodes: $\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \times \boldsymbol{\sigma}$, where $\boldsymbol{\epsilon}$ represents randomly generated noises.

The decoder part reconstructs the original graph by calculating the inner product between latent vectors:

$$p\left(A_{ij} = 1 \mid z_i, z_j\right) = \sigma\left(z_i^{\top} z_j\right), \qquad (3)$$

where A_{ij} is the element of the adjacency matrix of **A** and σ is the logistic sigmoid function.

The loss function during learning contains two parts, the first part measures the similarity between the constructed adjacency matrix and the input one, and the second part applies KL-divergence to measure how similar the distribution of the latent vectors and normal distribution is

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p\left(\mathbf{A} \mid \mathbf{Z}\right)\right] - KL\left[q\left(\mathbf{Z} \mid \mathbf{X}, \mathbf{A}\right) \| p\left(\mathbf{Z}\right)\right] \qquad (4)$$

By optimizing the loss function using gradient descent, we get the embedding matrix **Z** which represents the social homophily of social network users. The homophily representation for user u_i is denoted as z_{u_i}.

3.2.2 Social Group Representation

Previous studies [21–23] have shown that the more communities are affected in the early diffusion episodes of online content, the more viral it is. Since social network users usually only participate in the diffusion of topics they are interested in, users spontaneously form different social groups according to their interests. In this section, we try to model the social group diversity of early infected users to predict the final popularity of online information.

For user historical text sequence $\mathcal{M} = \{m_1, m_2 \ldots, m_N\}$, we use the sentence transformer model [49] in the field of natural language processing to generate semantic embedding for each text. For user u_i, the historical text embeddings are represented as $\{\boldsymbol{m}_{s_0}, \boldsymbol{m}_{s_1}, \ldots, \boldsymbol{m}_{s_k}\}, \boldsymbol{m}_{s_i} \in \mathbb{R}^L$, where L is the dimension of semantic embedding. Online content that users have participated in the past reflects the user's interest. Therefore, we construct the embedding of the user's interest based on the embeddings of the user historical text sequence. Referring to the mean pooling strategy in the convolutional neural network, we get the user interest representation $\boldsymbol{m}_{u_i}$ for each user by averaging the embeddings of the user history text: $\boldsymbol{m}_{u_i} = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{m}_{s_j}$.

Depending on the difference in user interest, we cluster users into different groups using an unsupervised clustering method—KMeans. The KMeans algorithm aims to divide data points into k different cluster sets to minimize the sum of the distance between sample points in each cluster set and the cluster center. We use $\boldsymbol{M} = \{\boldsymbol{m}_{u_i} \mid u_i \in \mathcal{U}\}$ as the input to the KMeans algorithm, and get the set of cluster center: $\boldsymbol{C} = \{\boldsymbol{c}_j\}, j \in 1, 2, \ldots, k$.

We use the relative position of each user to different social groups as a representation of the user's social group characteristic, that is, which social group a user is more likely to belong to. For user u_i, we measured the distance between the user and each cluster center by calculating the cosine distance between them:

$$d_{u_i,j} = \boldsymbol{m}_{u_i} \cdot \boldsymbol{c}_j \qquad (5)$$

Therefore, the social group representation $\boldsymbol{d}_{u_i}$ can be calculated by a dot product of user interest representation and the cluster center set: $\boldsymbol{d}_{u_i} = \boldsymbol{m}_{u_i} \cdot \boldsymbol{C}$.

3.2.3 Social Influence Representation

Social influence usually refers to the phenomenon that people's emotions, opinions, and behaviors are influenced by others. Many studies have shown that the social influence of early infected users directly affects the final diffusion scale of online content. The global retweet network reflects the social influence relationship among users, and the social influences between users are directional. The challenge is that the social strengths between

users are heterogeneous. Therefore, the learning of the latent representation of social influence should consider social relation strengths. We incorporate an attention mechanism with graph neural networks to characterize user influence from the directed global retweet network.

A graph neural network generally generates the node representation by aggregating information from the neighbors of the node, as the following function shows:

$$h_{u_i} = \sigma\left(\mathbf{W} \cdot \text{Aggre}\left(\{h_v, \forall v \in N(u_i)\}\right) + b\right) \qquad (6)$$

where $N(u_i)$ denotes the neighbor set of user u_i, Aggre represents the graph aggregation function, and h_{u_i} is the social influence representation of user u_i. In many cases, we use the element-wise mean operation of the neighbor embeddings as the Aggre function. But considering the diversity of the strength of social relations, we introduce the attention weights α_{iv} to denote the social strength between node u_i and its neighbor v. In other words, the weight means the contribution of neighbor v to generate the social influence of user u_i using the following formula,

$$h_{u_i} = \sigma\left(\mathbf{W} \cdot \left\{\sum_{v \in N(u_i)} \alpha_{iv} h_v\right\} + b\right) \qquad (7)$$

We automatically get the attention weight with a two-layer neural network which is defined as

$$\alpha_{iv}^* = \mathbf{w}_2^T \cdot \sigma\left(\mathbf{W}_1 \cdot \text{Concat}\left(h_v, \boldsymbol{p}_{u_i}\right) + b_1\right) + b_2 \qquad (8)$$

where $\boldsymbol{p}_{u_i}$ is the initial embedding of user u_i, W and b are the weight and bias of the neural network, respectively, and σ denotes the nonlinear activation function like rectified linear unit (ReLU). The final attention weight is normalized through the Softmax function,

$$\alpha_{iv} = \frac{\exp\left(\alpha_{iv}^*\right)}{\sum_{v \in N(i)} \exp\left(\alpha_{iv}^*\right)} \qquad (9)$$

3.2.4 Predictor

This section attempts to predict the popularity of online content based on the social influence, homophily, and social group representation of early infected users.

We obtain the social influence representation h_u, homophily representation z_u, and social group representation d_u for users in previous sections. For each online information and its infected user sequence $D(t) = \{u_1, u_2, .., u_k\}$, we need to generate the integrated embedding from the user sequence which describes the characteristics of information diffusion.

There are many methods for the fusion of user sequence embeddings, such as the LSTM neural network to learn the temporal dependency of users in CasCN [11], co-attentive fusion strategy adopted in Inf-Vae [17], etc. But the attention mechanism induces large computational cost and has limited improvements on prediction accuracy; hence, we take the element-wise sum operation of the user sequence embeddings and use a concatenation of different features as the fusion method, and finally use a two-layer MLP neural network to generate the predicted popularity, which is defined as

TABLE 1 | Dataset statistics.

Datasets	Nodes	Edges	Avg. popularity	Texts	Samples
Weibo	35,084	74,767	18.15	57,744	6,374
Aminer	35,800	155,987	32.41	234,970	30,311

$$g_1 = \left[\sum_{u_i \in D(t)} h_{u_i} \| z_{u_i} \| d_{u_i}\right] \qquad (10)$$

$$g_2 = \sigma\left(\mathbf{W}_2 \cdot g_1 + b_2\right) \qquad (11)$$

$$\Delta\hat{P}\left(t_f\right) = \sigma\left(\mathbf{w}^T \cdot g_2 + b\right) \qquad (12)$$

where $\|$ denotes the concatenation operation between embeddings, W and b are the weight and bias of the neural network, respectively, σ denotes the nonlinear activation function, and $\Delta\hat{P}(t_f)$ is the predicted incremental popularity of online contents.

Following many previous works [11,39], we use the mean square log-transformed error (MSLE) as the loss function and take log-transformation for the popularity to avoid the situation where the training process is dominated by large information cascades:

$$\mathcal{L}\left(\Delta P, \Delta\hat{P}\right) = \frac{1}{N} \sum_{i=0}^{N-1} \left(\log_2 \Delta P_i - \log_2 \Delta\hat{P}_i\right)^2 \qquad (13)$$

where N is the total number of samples in our training dataset, $\Delta\hat{P}_i$ is the predicted incremental popularity of online information, and ΔP_i is the actual incremental popularity of online information. Model parameters are trained by minimizing the loss function and optimized using the Adam algorithm, given its efficiency and ability to avoid overfitting.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

We evaluate the performance of our proposed model for the popularity prediction task on two real datasets. Sina Weibo is a popular Twitter-like social network platform in China which allows users to post user-generated contents or retweet messages they are interested in from other users. We use the keywords related to some real social events, such as "Baidu suing Bytedance," and "D.Y discharge hazardous waste," to crawl the corresponding diffusion process of the microblogs on Sina Weibo from April–July 2021 as one of our datasets. We use this dataset to predict the incremental popularity of microblogs. In addition, we use the publicly available article citation dataset published on Aminer [50], which is mainly used for the extraction and mining of academic social networks. Each academic article and its related citation form a diffusion cascade and we take the abstract of the article as its corresponding semantic information. We use this dataset to predict the citation increment of articles.

To prevent information leakage, we extract the diffusion data before the observation time from the crawled data. Similar to many existing works [11], we eliminate the microblogs with more

than 1,000 retweets because the number of large cascades is very small and may dominate the training process. The detailed statistics of datasets are shown in **Table 1**.

Following previous works [11,51], we adopt the mean square log-transformed error (MSLE) and mean absolute error (MAE) as our evaluation metrics to measure the difference between the actual value and the predicted value, which are defined as

$$MSLE = \frac{1}{N} \sum_{i=0}^{N-1} \left(\log_2 \Delta P_i - \log_2 \Delta \hat{P}_i \right)^2, \qquad (14)$$

$$MAE = \frac{1}{N} \sum_{i=0}^{N-1} |\log_2 \Delta P_i - \log_2 \Delta \hat{P}_i|, \qquad (15)$$

where N is the total number of samples, $\Delta \hat{P}_i$ is the predicted incremental popularity of online information, and ΔP_i is the actual incremental popularity of online information.

4.2 Baselines

To evaluate the effectiveness of our model in popularity prediction, we make an experimental comparison between our model and the following methods. The baselines include a traditional feature-based method SH and three classical deep learning-based methods: DeepWalk, CasCN, and CasFlow. These methods either ignore the modeling of homophily or overlook the directivity of social influence when modeling the social influence of users:

(1) The SH model [24] is the most widely used feature-based popularity prediction model, which uses the strong linear correlation between the early popularity and the final popularity.
(2) DeepWalk [52] learns low-dimensional node embedding by taking random walks on social networks and using the classical Word2Vec method to model the sampled node sequences. We use the DeepWalk method to learn social influence representation from global retweet networks and apply the learned social influence representation of early adopters for popularity prediction tasks.
(3) CasCN [11] is a classical deep learning-based popularity prediction model. It proves the effectiveness of using a graph neural network framework to generate the cascade graph representation for popularity prediction tasks. By learning the cascade subgraph representation using the GNN neural network and modeling the dynamic change of the cascade graph using the LSTM neural network, the method uses both the temporal and structural information in the early information diffusion episode.
(4) CasFlow [13] is the latest approach on popularity prediction. It is an extension of the VaCas model [12]. It uses GraphWave to learn the influence representation of nodes in cascade graphs and uses sparse matrix factorization to learn the user preference representation in global retweet graphs.

We use grid search to set the super-parameters. For a Weibo dataset, the batch size is 8, the learning rate is 0.003, and we set the weight decay to be 0.002 to prevent model overfitting. For an

Aminer dataset, the batch size is 16, the learning rate is 0.003, and the weight decay is 0.002. The number of social groups is 8 for both Weibo and Aminer datasets. For the aforementioned two datasets, we use the 10-fold cross-validation method and take the average of results of each fold to evaluate the performance of the proposed model.

We use SentenceTransformers [49] to generate the semantic embedding of texts and the recommended KMeans clustering method on the official website of SentenceTransformers. The generated embeddings can be compared using cosine-similarity to find texts with similar meanings. The amount of user-related texts in the initial datasets is quite large as shown in **Table 1**. We use the pre-trained models—a multilingual knowledge distilled version which supports Chinese languages for the Weibo dataset and the pre-trained model—paraphrase MiniLM for the Aminer dataset.

4.3 Variants

We provide three variants of the proposed model to investigate and demonstrate the validity of each component. Variants are generated by removing some parts of our proposed model.

(1) Inf-HG-NI: We remove the social influence component from the model and use the group and social homophily representation of early infected users for popularity prediction.
(2) Inf-HG-NG: We remove the social group component from the model and use the social influence and homophily representation of early infected users for popularity prediction.
(3) Inf-HG-NH: We remove the homophily component from the model and use the social influence and social group representation of early infected users for popularity prediction.

4.4 Experiment Results

The experimental results are shown in **Table 2**, where T refers to the observation time window. We set the confidence level to 0.95 and carry out a paired t-test between the best baseline method and our proposed model, and the calculated significance level p-value is less than 0.05. The experimental results of the proposed method on the two datasets show that the proposed model has better prediction results than the baseline methods, surpassing the sub-optimal model by around 5%.

4.4.1 Performance Comparison

The prediction accuracy of the traditional feature-based SH model is poorer. The model is quite simple and only uses the historical popularity information before the observation time which leads to less satisfying prediction results. The DeepWalk method can learn the social influence representation of users who participate in information diffusion, but it only uses the social influence of users. Limited by its representation ability, the overall performance of this model is not quite good. But the model does not require too many complicated data preprocessing operations and can be easily implemented. CasCN innovatively uses the latest graph neural network framework to model the cascade

TABLE 2 | Overall performance comparison.

Datasets	Weibo				Aminer			
T	1 hour		2 hours		3 Years		5 Years	
Metric	MSLE	MAE	MSLE	MAE	MSLE	MAE	MSLE	MAE
SH	2.420	1.190	2.315	1.186	1.358	0.884	1.324	0.821
DeepWalk	2.287	1.128	2.199	1.127	1.148	0.788	1.139	0.763
CasCN	2.134	1.122	2.106	1.131	1.003	0.712	1.011	0.723
CasFlow	1.891	1.003	1.680	0.997	0.926	0.725	1.014	0.758
Inf-HG	1.929	1.097	1.629	0.995	0.846	0.693	0.839	0.702
p-value	< 0.021	< 0.004	< 0.016	< 0.030	< 0.001	< 0.001	< 0.001	< 0.001
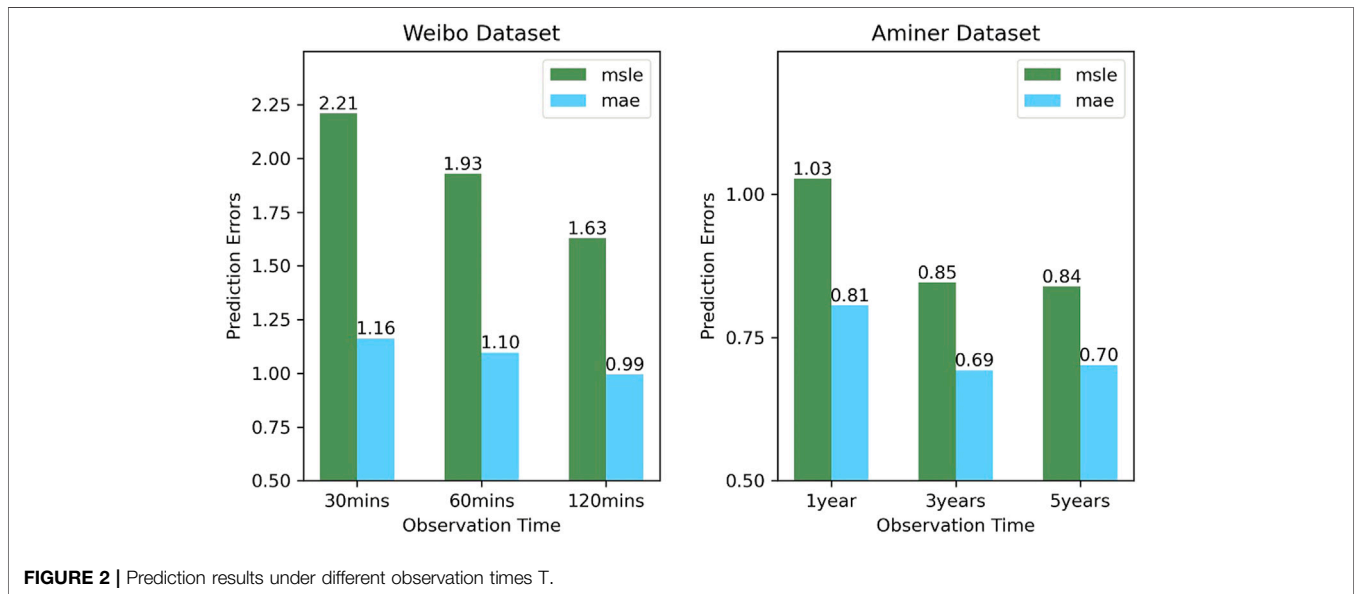
FIGURE 2 | Prediction results under different observation times T.

graph at the graph-level and also pay attention to capture the evolution of the cascade graph over time. The prediction accuracy of CasCN is significantly improved compared with the DeepWalk method. However, this method mainly models the social influence of users, while overlooking the importance of homophily and the semantic information of online contents. Therefore, there is still room for further improvement in prediction accuracy. In addition, the model is complicated and requires quite a long training time.

CasFlow [13] is the latest approach to popularity prediction and achieved better prediction results than other baselines. It uses sparse matrix factorization to learn the user preference representation in global retweet graphs and the GraphWave method which is an unsupervised spectral graph wavelet method to model the social influence of users in the cascade graph. However, this model does not consider the direction of user influence because the Laplacian matrix cannot be applied to directed graphs, thus failing to effectively model the social influence of users.

Compared with the CasFlow model, our model shows comparable or better performance on dataset Weibo and has a significant performance improvement on dataset Aminer. Our approach uses the attention-based graph neural network to model

the directed social influence of early adopters and proposes a novel method to jointly model the social influence and homophily when the underlying network structure is unknown. The performance gain over CasFlow illustrates the superiority of modeling the social influence, homophily, and social group information for popularity prediction.

On the other hand, we can see from the table that the prediction errors of both MSLE and MAE decrease as the observation time window becomes large. We conduct more experiments on our model, and the results are shown in **Figure 2**. We can see an obvious decline in the prediction errors over time. The reason is that we are allowed to observe more diffusion of information; moreover, on many social network platforms, newly posted information tends to gain more attention, whereas old information quickly loses attention as time goes by and its popularity tends to stay the same. Hence, as the observation time increases, the diffusion process is more deterministic, which makes the prediction easier.

4.4.2 Variant Comparison

The performance of three variants is shown in **Table 3**, where we compare the results to investigate and demonstrate the validity of each component. The p-value is calculated between each variant

TABLE 3 | Variant comparison.

Datasets	Weibo				Aminer			
Metric	MSLE	*p*-value	MAE	*p*-value	MSLE	*p*-value	MAE	*p*-value
Inf-HG-NI	2.007	< 0.001	1.133	< 0.001	0.929	< 0.001	0.715	< 0.001
Inf-HG-NG	1.934	< 0.022	1.091	< 0.025	0.870	< 0.001	0.705	< 0.003
Inf-HG-NH	1.981	< 0.009	1.119	< 0.018	0.895	< 0.001	0.706	< 0.003
Inf-HG	1.929	–	1.097	–	0.846	–	0.693	–

FIGURE 3 | Prediction results under different cluster numbers.
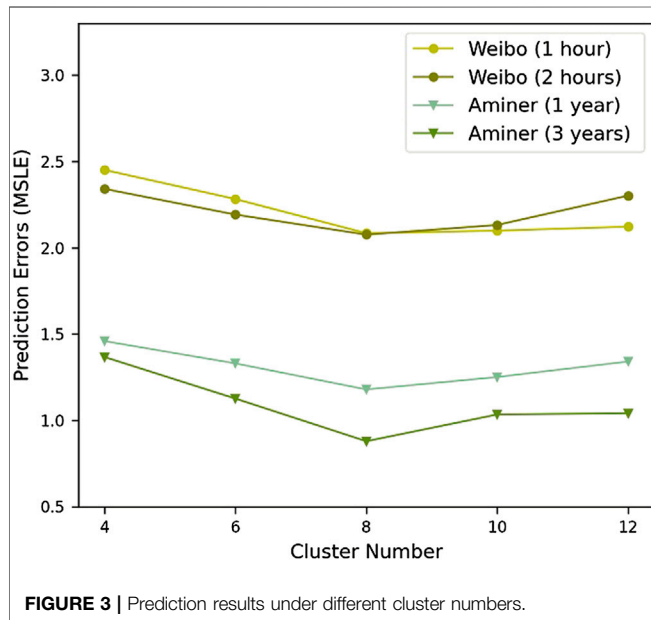
and our proposed model. The observation time window for the Weibo and Aminer datasets is 1 h and 3 years, respectively. Each variant performed worse than the proposed model, suggesting that modeling multiple features simultaneously can significantly improve the prediction accuracy.

The prediction result obtained by removing the social group characteristics of users (Inf-HG-NG) is worse than the proposed model, which proves the effectiveness of clustering users into different social groups based on their interests and applying the social group characteristics of users to make the prediction of information diffusion. The experimental results are consistent with traditional feature-based research [21,23], which demonstrates that the future popularity of online information can be predicted by modeling its early spreading pattern in the social community. It is to be noted that we generate the social group representation based on the user interest with a social network unknown which is different from previous studies.

The variant Inf-HG-NG uses the social influence and homophily representation for prediction and performs better than the other two variants, which can also demonstrate that the importance of jointly modeling the social influence and homophily of the diffusion episodes in popularity predictions.

The prediction result obtained by removing the social influence characteristics (Inf-HG-NI) performs the worst on both datasets, which suggests that social influence shows more importance compared with social homophily and social group characteristics, and also confirms that our model can better capture the social influence of users by using the attention-based graph neural network framework.

We also show the prediction results under different cluster numbers in **Figure 3**. We generate different cluster center sets according to the cluster number and use the social group representation of early infected users as the input to the two-layered neural network for prediction. The number of social groups for both datasets is 8 to achieve the best performance.

5 CONCLUSION

In this study, we attempt to model the joint effect of social influence, homophily, and social group features of early infected users for predicting the popularity of online contents in the case where the social network is unknown. We use two different graph neural networks to model social homophily and social influence. By integrating an attention mechanism with a graph neural network, we can effectively learn the heterogeneous social strength to generate the social influence representation. We also attempt to generate social group representation by clustering users into different groups based on user interest. Experimental results on real-world datasets demonstrate the effectiveness of our proposed method; moreover, the enhancement of prediction accuracy confirms the validity of jointly modeling the social influence, homophily, and social group characteristics of early infected users in popularity prediction tasks. In the future, we will consider other methods to model the effect of social groups on the popularity of online content.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YS contributed to the core idea of the experiment design and wrote the manuscript under the guidance of BZ. BZ supervised

FUNDING

REFERENCES

1. Gao X, Cao Z, Li S, Yao B, Chen G, Tang S. Taxonomy and Evaluation for Microblog Popularity Prediction. *ACM Trans Knowledge Discov Data* (2019) 13. doi:10.1145/3301303

2. Zhou F, Xu X, Trajcevski G, Zhang K. A Survey of Information Cascade Analysis: Models, Predictions and Recent Advances. *ACM Comput Surv* (2020) 1:1–41.

3. Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, et al. A 61-Million-Person experiment in Social Influence and Political Mobilization. *Nature* (2012) 489:295–8. doi:10.1038/nature11421

4. Wu Q, Gao Y, Gao X, Weng P, Chen G. Dual Sequential Prediction Models Linking Sequential Recommendation and Information Dissemination. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019). p. 447–57. doi:10.1145/3292500.3330959

5. Leskovec J, Adamic LA, Huberman BA. The Dynamics of Viral Marketing. *ACM Trans Web* (2007) 1:5. doi:10.1145/1232722.1232727

6. Cheng J, Adamic LA, Dow PA, Kleinberg J, Leskovec J. Can Cascades Be Predicted? In: WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web (2014). p. 925–35.

7. Shulman B, Sharma A, Cosley D. Predictability of Popularity: Gaps between Prediction and Understanding. In: Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016 (2016). p. 348–57.

8. Bakshy E, Mason WA, Hofman JM, Watts DJ. Everyone's an Influencer: Quantifying Influence on Twitter. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011 (2011). p. 65–74.

9. Shen H, Wang D, Song C, Barabási AL. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. *Proc Natl Conf Artif Intelligence* (2014) 1:291–7.

10. Li C, Ma J, Guo X, Mei Q. DeepCas: An End-To-End Predictor of Information Cascades. In: 26th International World Wide Web Conference, WWW 2017 (2017). p. 577–86.

11. Chen X, Zhou F, Zhang K, Trajcevski G, Zhong T, Zhang F. Information Diffusion Prediction via Recurrent Cascades Convolution. In: Proceedings - International Conference on Data Engineering 2019-April (2019). p. 770–81. doi:10.1109/icde.2019.00074

12. Zhou F, Xu X, Zhang K, Trajcevski G, Zhong T. Variational Information Diffusion for Probabilistic Cascades Prediction. In: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications (2020). p. 3. doi:10.1109/infocom41043.2020.9155349

13. Xu X, Zhou F, Zhang K, Liu S, Trajcevski G. CasFlow: Exploring Hierarchical Structures and Propagation Uncertainty for Cascade Prediction. *IEEE Trans Knowledge Data Eng* (2021) 4347:1–15.

14. Zhang J, Dong Y, Wang Y, Tang J, Ding M. Prone: Fast and Scalable Network Representation Learning. In: IJCAI-19: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization (2019). p. 4278–84. doi:10.24963/ijcai.2019/594

15. Aral S, Muchnik L, Sundararajan A. Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks. *Proc Natl Acad Sci U.S.A* (2009) 106:21544–9. doi:10.1073/pnas.0908800106

16. Shalizi CR, Thomas AC. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods Res* (2011) 40:211–39. doi:10.1177/0049124111404820

17. Sankar A, Zhang X, Krishnan A, Han J. InF-VAE: A Variational Autoencoder Framework to Integrate Homophily and Influence in Diffusion Prediction. In: WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining (2020). p. 510–8.

18. Jenders M, Kasneci G, Naumann F. Analyzing and Predicting Viral Tweets. In: WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web (2013). p. 657–64. doi:10.1145/2487788.2488017

19. Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J. DeepInf: Social Influence Prediction with Deep Learning. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018). p. 2110–9.

20. Kipf TN, Welling M. Variational Graph Auto-Encoders. In: NIPS-W (2016). p. 1–3.

21. Weng L, Menczer F, Ahn YY. Virality Prediction and Community Structure in Social Networks. *Sci Rep* (2013) 3:2522. doi:10.1038/srep02522

22. Weng L, Menczer F, Ahn YY. Predicting Successful Memes Using Network and Community Structure. In: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014 (2014). p. 535–44.

23. Liu C, Wang W, Sun Y. Community Structure Enhanced cascade Prediction. *Neurocomputing* (2019) 359:276–84. doi:10.1016/j.neucom.2019.05.069

24. Szabo G, Huberman BA. Predicting the Popularity of Online Content. *Commun ACM* (2011) 53(8). doi:10.2139/ssrn.1295610

25. Pinto H, Almeida JM, Gonçalves MA. Using Early View Patterns to Predict the Popularity of YouTube Videos. In: WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining (2013). p. 365–74. doi:10.1145/2433396.2433443

26. Bao P, Shen HW, Huang J, Cheng XQ. Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo. In: WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web (2013). p. 177–8.

27. Tsur O, Rappoport A. What's in a Hashtag? Content Based Prediction of the Spread of Ideas in Microblogging Communities. In: WSDM 2012- Proceedings of the 14th ACM International Conference on Web Search and Data Mining (2012). p. 643–52.

28. Ma Z. On Predicting the Popularity of Newly Emerging Hashtags in Twitter. *J Am Soc Inf Sci Technol* (2013) 64:1852–63. doi:10.1002/asi.22844

29. Pastor-Satorras R, Vespignani A. Epidemic Spreading in Scale-free Networks. *Phys Rev Lett* (2001) 86:3200–3. doi:10.1103/physrevlett.86.3200

30. Lin S, Kong X, Yu PS. Predicting Trends in Social Networks via Dynamic Activeness Model. In: International Conference on Information and Knowledge Management, Proceedings (2013). p. 1661–6. doi:10.1145/2505515.2505607

31. Shen H, Wang D, Song C, Barabási A-L. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In: AAAI'14: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI Press) (2014). p. 291–7.

32. Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J. SEISMIC: A Self-Exciting point Process Model for Predicting Tweet Popularity. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2015-Augus (2015). p. 1513–22.

33. Gao L, Liu Y, Zhuang H, Wang H, Zhou B, Li A. Public Opinion Early Warning Agent Model: A Deep Learning Cascade Virality Prediction Model Based on Multi-Feature Fusion. *Front Neurorobotics* (2021) 15:1–12. doi:10.3389/fnbot.2021.674322

34. Gao L, Wang H, Zhang Z, Zhuang H, Zhou B. Hetinf: Social Influence Prediction with Heterogeneous Graph Neural Network. *Front Phys* (2022) 9. doi:10.3389/fphy.2021.787185

35. Shang Y, Zhou B, Wang Y, Li A, Chen K, Song Y, et al. Popularity Prediction of Online Contents via cascade Graph and Temporal Information. *Axioms* (2021) 10. doi:10.3390/axioms10030159

36. Kwak H, Lee C, Park H, Moon S. What Is Twitter, a Social Network or a News media? In: Proceedings of the 19th International Conference on World Wide Web (2010). p. 591–600. Vol. 19.

37. Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring User Influence in Twitter: The Million Follower Fallacy. In: AAAI Conference on Weblogs and Social Media (2010). p. 10–7. Vol. 14.

38. Bao P. Modeling and Predicting Popularity Dynamics via an Influence-Based Self-Excited Hawkes Process. In: CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery (2016). p. 1897–900. doi:10.1145/2983323.2983868

39. Cao Q, Shen H, Cen K, Ouyang W, Cheng X. DeepHawkes: Bridging the gap between Prediction and Understanding of Information Cascades. In: International Conference on Information and Knowledge Management, Proceedings Part F1318 (2017). p. 1149–58.

40. Yang C, Tang J, Sun M, Cui G, Liu Z. Multi-scale Information Diffusion Prediction with Reinforced Recurrent Networks. In: IJCAI International Joint Conference on Artificial Intelligence 2019-Augus (2019). p. 4033–9. doi:10.24963/ijcai.2019/560

41. Cao Q, Shen H, Gao J, Wei B, Cheng X. Popularity Prediction on Social Platforms with Coupled Graph Neural Networks. In: WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining (2020). p. 70–8. doi:10.1145/3336191.3371834

42. Liu Z, Wang R, Liu Y. Prediction Model for Non-topological Event Propagation in Social Networks. *Commun Comput Inf Sci* (2019) 1058: 241–52. doi:10.1007/978-981-15-0118-0_19

43. Newman MEJ. Modularity and Community Structure in Networks. *Proc Natl Acad Sci U.S.A* (2006) 103:8577–82. doi:10.1073/pnas.0601602103

44. Nematzadeh A, Ferrara E, Flammini A, Ahn YY. Optimal Network Modularity for Information Diffusion. *Phys Rev Lett* (2014) 113:088701. doi:10.1103/PhysRevLett.113.088701

45. Al-Taie MZ, Kadry S. Modeling Information Diffusion in Social Networks Using Latent Topic Information. *Adv Inf Knowledge Process* (2014) 2014:165–84.

46. Zhang Q, Gong Y, Wu J, Huang H, Huang X. Retweet Prediction with Attention-Based Deep Neural Network. In: International Conference on Information and Knowledge Management, Proceedings 24-28-Octo (2016). p. 75–84. doi:10.1145/2983323.2983809

47. Wang S, Zhou L, Kong B. Information cascade Prediction Based on T-DeepHawkes Model. *IOP Conf Ser Mater Sci Eng* (2020) 715. doi:10.1088/1757-899x/715/1/012042

48. Wang H, Yang C, Shi C. Neural Information Diffusion Prediction with Topic-Aware Attention Network. In: International Conference on Information and Knowledge Management, Proceedings (2021). p. 1899–908. doi:10.1145/3459637.3482374

49. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics (2019). p. 3982–92. doi:10.18653/v1/d19-1410

50. Tang J, Zhang J. Arnetminer: Extraction and Mining of Academic Social Network. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008). p. 990–8.

51. Liu Y, Bao Z, Zhang Z, Tang D, Xiong F. Information Cascades Prediction with Attention Neural Network. *Human-centric Comput Inf Sci* (2020) 10. doi:10.1186/s13673-020-00218-w

52. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014). p. 701–10.