

Continuous-Time Graph Learning for Cascade Popularity Prediction

Xiaodong Lu, Shuo Ji, Le Yu, Leilei Sun, Bowen Du, Tongyu Zhu*

SKLSDE Lab, Beihang University, Beijing 100191, China

{xiaodonglu, jishuo, yule, leileisun, dubowen, zhutongyu}@buaa.edu.cn

Abstract

Information propagation on social networks could be modeled as cascades, and many efforts have been made to predict the future popularity of cascades. However, most of the existing research treats a cascade as an individual sequence. Actually, the cascades might be correlated with each other due to the shared users or similar topics. Moreover, the preferences of users and semantics of a cascade are usually continuously evolving over time. In this paper, we propose a continuous-time graph learning method for cascade popularity prediction, which first connects different cascades via a universal sequence of user-cascade and user-user interactions and then chronologically learns on the sequence by maintaining the dynamic states of users and cascades. Specifically, for each interaction, we present an evolution learning module to continuously update the dynamic states of the related users and cascade based on their currently encoded messages and previous dynamic states. We also devise a cascade representation learning component to embed the temporal information and structural information carried by the cascade. Experiments on real-world datasets demonstrate the superiority and rationality of our approach.

1 Introduction

The information propagation, aka the information cascade, is ubiquitous on online social networks, which records human behaviors in posting and accessing information. For example, on Twitter, a tweet posted by a user may disseminate to other users, and such retweeting behaviors between users can be denoted as an information cascade. Predicting the popularity of such information cascades could help people understand the information propagation better and is crucial for numerous applications such as viral marketing [Leskovec *et al.*, 2007], scientific impact qualification [Guo and Suo, 2014] and item recommendation [Wu *et al.*, 2019].

Up to now, lots of attempts have been made on this problem. In the early stage, researchers extracted man-

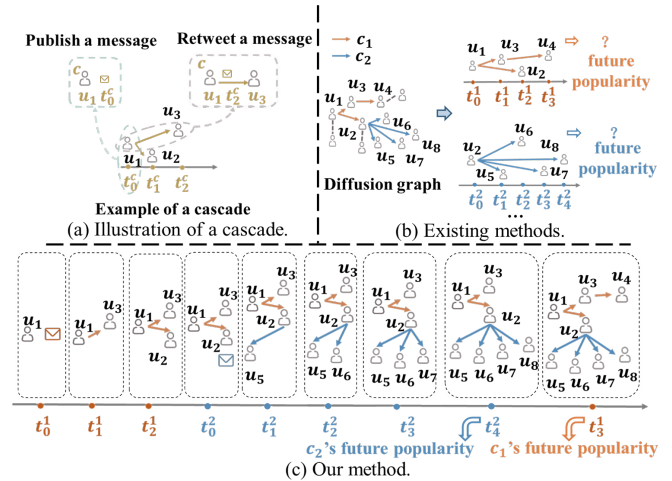


Figure 1: Different from existing methods which mainly learn on each cascade’s own sequence, our method learns continuously evolving representations of cascades and users collaboratively, which can model the correlation between cascades and continuously dynamic user preferences.

ual features to represent a cascade [Cheng *et al.*, 2014; Szabó and Huberman, 2010]. However, these methods are based on hand-designed features with a relatively large number of human efforts and may lose useful information during the diffusion behavior of a cascade. Different from feature-based methods, some researchers considered the cascade as a diffusion sequence and employed sequential models like recurrent neural networks to capture the evolution pattern of cascades [Liao *et al.*, 2019; Cao *et al.*, 2017; Yang *et al.*, 2019], while the structural information within the cascade has not been fully exploited yet. Recently, graph representation learning methods were introduced to further improve the prediction performance [Li *et al.*, 2017; Chen *et al.*, 2019b; Chen *et al.*, 2019a; Xu *et al.*, 2021; Tang *et al.*, 2021]. These methods utilized the social network and cascade graph to learn the structural and temporal information within each cascade.

Although insightful, most of the existing methods solely predict the popularity of each cascade within its own sequence, see Figure 1 (b). We argue that there are two essen-

*Corresponding Author

tial factors that are not well considered by previous methods. Firstly, different cascades can be correlated with each other because of the shared users or similar semantics. For example, in Figure 1 (b), we aim to predict the popularity of cascade c_1 and existing methods can only learn from the propagation sequence of c_1 itself. However, if we additionally consider cascade c_2 , we can find that both c_1 and c_2 contain user u_2 which seems to be a popular user with many retweets, and this is helpful for predicting the popularity of c_1 . Secondly, the states of users are often evolving in a continuous manner (e.g., a user is likely to gradually change his interests according to the information he/she received from the social network at different times), which cannot be captured by existing methods either.

To tackle the above issues, we propose a Continuous-Time graph learning method for Cascade Popularity prediction, namely CTCP. To model the correlation between cascades, we first combine all cascades into a dynamic diffusion graph as shown in Figure 1 (c), which can be considered as a universal sequence of diffusion behaviors (i.e, user-cascade and user-user interactions). Then, we propose an evolution learning module to chronologically learn on each diffusion behavior by maintaining a dynamic representation for each user and cascade that evolves continuously as the diffusion behavior happens. When a diffusion behavior happens, this module first encodes the information of a diffusion behavior into a message and then fuses the dynamic representations of related users and cascades with the generated message. Next, a cascade representation learning model is proposed to generate the static and dynamic cascade embeddings by aggregating user representations from both temporal and structural perspectives. Based on the generated embeddings, the prediction module finally computes the cascade popularity. The main contributions of the paper are summarized as follows.

- Different from previous methods that only learn from the own sequence of each cascade, we propose a continuous-time graph learning method to explore the correlations of different cascades by a dynamic diffusion graph and explicitly learn the dynamic preferences of users in the network.
- We maintain dynamic representations for users and cascades and design an evolution learning module to encode the information of a diffusion behavior into a message and continuously update the dynamic representations by fusing the previous dynamic representations with the message in a recurrent manner.
- A cascade representation learning module is proposed to capture the temporal and structural information of a cascade, which leverages the sequence and graph structure to aggregate representations of users.

2 Related Work

2.1 Cascade Popularity Prediction

The cascade popularity prediction problem aims at predicting the future size of an information cascade. Many efforts have been paid to this problem. In the early stage, researchers represented a cascade as some handcrafted features such

as content features and user attributes [Cheng *et al.*, 2014; Szabó and Huberman, 2010]. However, these methods need a relatively large number of human labor to design or select and have limited generalization ability.

Different from the feature-based methods, some researchers considered the cascade as a diffusion sequence of users and employed the sequence-based model to learn the evolution pattern of a cascade [Liao *et al.*, 2019; Cao *et al.*, 2017; Yang *et al.*, 2019]. For example, Cao *et al.*[2017] utilized the Gated Recurrent Unit (GRU) to learn a path-level representation and aggregate path representation into cascade representation by different learnable weights. Though the sequence-based methods achieve considerable performance, the structural information of cascades is not well explored.

To fully utilize the temporal and structural information within cascades, some graph-based methods have been proposed [Li *et al.*, 2017; Chen *et al.*, 2019b; Chen *et al.*, 2019a; Xu *et al.*, 2021; Tang *et al.*, 2021], which modeled a single cascade as a graph evolving with time and leveraged the graph representation learning method to learn cascade representations from cascade graphs. However, these methods predict the popularity of each cascade separately and thus neglect the correlation between cascades. Some recent methods model the evolution of multiple cascades by considering it as a sequence of graph snapshots sampled at regularly-spaced times [Wang *et al.*, 2021; Sun *et al.*, 2022], but these methods discretize the continuous timestamps into several regularly-spaced time steps and thus can not model the continuous evolution of user preferences. Moreover, these methods may have high memory overhead, because it needs to load a whole graph snapshot at one time.

In summary, although insightful, existing methods have not well addressed the issues of correlation between cascades and the dynamic evolution of user preferences.

2.2 Graph Representation Learning

In recent years, the Graph Neural Network (GNN) has achieved superior performance on graph representation learning. To be better in accordance with real-world scenarios, some researchers have further designed heterogeneous GNNs and dynamic GNNs [Kazemi *et al.*, 2020; Wang *et al.*, 2020; Huang *et al.*, 2021]. For example, Schlichtkrull *et al.*[2018] focused on the relation learning in knowledge graphs. Wang *et al.*[2019] and Hu *et al.*[2020] studied heterogeneous graphs based on meta-paths and the attention mechanism. Some researchers [Pareja *et al.*, 2020; Sankar *et al.*, 2020] treated a dynamic graph as a sequence of snapshots, while others [Xu *et al.*, 2020; Chang *et al.*, 2020; Rossi *et al.*, 2020] modeled each dynamic graph as a temporal graph or a sequence of events.

In this paper, we investigate the correlation between cascades and the dynamic user preferences by considering the evolution of cascades as a continuous-time graph.

3 Problem Formulation

Cascade. Given a set of users \mathcal{U} , a cascade c records the diffusion process of a message m among the users \mathcal{U} . Specifically, we use a chronological sequence $g^c(t) =$

$\{(u_i^c, v_i^c, t_i^c)\}_{i=1, \dots, |g^c(t)|}$ to represent the growth process of cascade c until time t , where (u_i^c, v_i^c, t_i^c) indicates that v_i^c forwards the message m from u_i^c (or we can say that v_i^c participates in cascade c through u_i^c). In addition, we use (u_0^c, t_0^c) to denote that u_0^c publishes the message m at t_0^c (or we can say that u_0^c begins cascade c at t_0^c).

Diffusion Graph. Based on the above definitions, we use the diffusion graph $\mathcal{G}_d^t = \{(u_i, v_i, c_i, t_i) | t_i < t\}$ to denote the diffusion process of all cascades until t . Here (u_i, v_i, c_i, t_i) is a diffusion behavior representing that v_i participates in cascade c_i through u_i at t_i . The diffusion graph \mathcal{G}_d^t can be considered as a chronological sequence of diffusion behaviors as shown in Figure 1 (c).

Cascade Prediction. Given a cascade c begins at t_0^c , after observing it for time t_o , we want to predict its incremental popularity $\Delta P_c = |g^c(t_0^c + t_p)| - |g^c(t_0^c + t_o)|$ from $t_0^c + t_o$ to $t_0^c + t_p$, where $t_p \gg t_o$ is the prediction time.

Most of the previous methods consider the task as a single cascade prediction problem, that is, learning a function $f : g^c(t_0^c + t_o) \rightarrow \Delta P_c$ that predicts the incremental popularity of a cascade only based on its own historical observation. However, the collaborative signals between the cascades are ignored, which motivates us to design our new method to consider other cascades when predicting the incremental popularity of a cascade. Specifically, we learn a function $f : g^c(t_0^c + t_o) \times G_d^{t_0^c + t_o} \rightarrow \Delta P_c$ which not only considers the information of a single cascade but also takes the historical diffusion on the social network into account.

4 Methodology

As shown in Figure 2, we first consider all cascades into a chronological sequence of diffusion behaviors (i.e., the diffusion graph). Then we learn on each diffusion behavior sequentially, where we maintain continuously evolving representations for cascades and users to explore the dynamic preference of users and the correlation between cascades. During the sequential learning process, whenever the observation time $t_o + t_0^c$ of a cascade c is reached, we predict its incremental popularity ΔP_c .

Specifically, our method consists of three components: 1) Evolution learning module maintains dynamic states (i.e., the dynamic representation) for users and cascades, which models cascades in diffusion behavior level (micro). 2) Cascade representation learning module generates the embeddings of cascades by aggregating user representations from different perspectives, which models cascades in a diffusion structure level (macro). 3) Prediction module gives the prediction of the incremental popularity of cascades.

4.1 Evolution Learning Module

Dynamic States. We first introduce the dynamic states for users and cascades. From the perspective of information diffusion, there are two roles of a user: originator and receiver. For example, in a diffusion behavior (u, v, c, t) , user u acts as the originator of the message, and user v acts as the receiver of the message. Thus, we maintain two types of dynamic states $s_u^o(t)$ and $s_u^r(t)$ for a user to describe his originator role and receiver role respectively. Besides, we main-

tain a dynamic state $s_c(t)$ for every cascade c to memorize its diffusion history, which can help users get information from previous participating users in the cascade. The above dynamic states are initialized to zero vectors and learned from the global diffusion behavior sequence.

Dynamic State Learning. When a diffusion behavior (u, v, c, t) happens, the dynamic states of the corresponding users and cascade should be updated. Naturally, the behaviors of a user (cascade) can be considered as a sequence and sequential models like the recurrent neural network can be employed to learn dynamic states from the sequence of a user (cascade). In addition to the own behaviors of a user (cascade), there are also global dependencies needed to be considered. For example, when a user u participates in a diffusion behavior (u, v, c, t) , he may also be influenced by the users who previously participated in the cascade c . To this end, we employ a recurrent neural network $f_r(\cdot)$ to update the dynamic states of users and cascades globally. Specifically, when a diffusion behavior (u, v, c, t) happens, we update the states of u, v, c by $f_r(\cdot)$. The updating process consists of two steps: interaction encoding and state updating. In the interaction encoding, we encode the information of diffusion behavior (u, v, c, t) and generate message $\mathbf{m}_u(t)$, $\mathbf{m}_v(t)$, $\mathbf{m}_c(t)$ for u, v and c to guide the subsequent state updating process. Assuming the state of u before t is $s_u^o(t^-)$, we generate message representation for user u by the following mechanism,

$$\mathbf{f}_u^t = [\cos w_1^r \Delta t_u, \cos w_2^r \Delta t_u, \dots, \cos w_n^r \Delta t_u], \quad (1)$$

$$\mathbf{m}_u(t) = \sigma(\mathbf{W}^r [s_u^o(t^-) || \mathbf{s}_v^r(t^-) || \mathbf{s}_c(t^-) || \mathbf{f}_u^t] + \mathbf{b}^r), \quad (2)$$

where $||$ is the concatenation operation, Δt_u is the time interval since the last updating of users u (i.e., $\Delta t_u = t - t_u^-$ and t_u^- is the last time where u was updated), and \mathbf{f}_u^t is the temporal feature learned from a series of cosine basis functions.

After generating the message representation, we fuse the old dynamic state $s_u^o(t^-)$ with the message representation $\mathbf{m}_u(t)$ to get the updated states $s_u^o(t)$ by GRU [Cho *et al.*, 2014],

$$\begin{aligned} \mathbf{g}_i &= \sigma(\mathbf{W}_{i,s} s_u^o(t^-) + \mathbf{W}_{i,m} \mathbf{m}_u(t) + \mathbf{b}_i), \\ \mathbf{g}_f &= \sigma(\mathbf{W}_{f,s} s_u^o(t^-) + \mathbf{W}_{f,m} \mathbf{m}_u(t) + \mathbf{b}_f), \\ \hat{s}_u^o(t) &= \tanh(\mathbf{W}_m \mathbf{m}_u(t) + \mathbf{g}_i \odot (\mathbf{W}_s s_u^o(t^-) + \mathbf{b}_s) + \mathbf{b}), \\ s_u^o(t) &= \mathbf{g}_f \odot s_u^o(t) + (1 - \mathbf{g}_f) \odot \hat{s}_u^o(t^-), \end{aligned} \quad (3)$$

The updating process of user v and cascade c is the same as user u in addition to different learnable parameters.

4.2 Cascade Representation Learning Module

In this module, we generate embeddings for cascades by aggregating representations of participating users. Specifically, we learn the temporal and structural characteristics of a cascade by leveraging the diffusion sequence and cascade graph to aggregate representations of users respectively. Besides the dynamic states $s_u^o(t)$ and $s_u^r(t)$ of users, we also introduce the static state s_u to represent the static preference of a user u . The static state is initialized randomly and learnable during the training process.

Temporal Learning. Given a cascade c , we organize it as a diffusion sequence of users $U_c =$

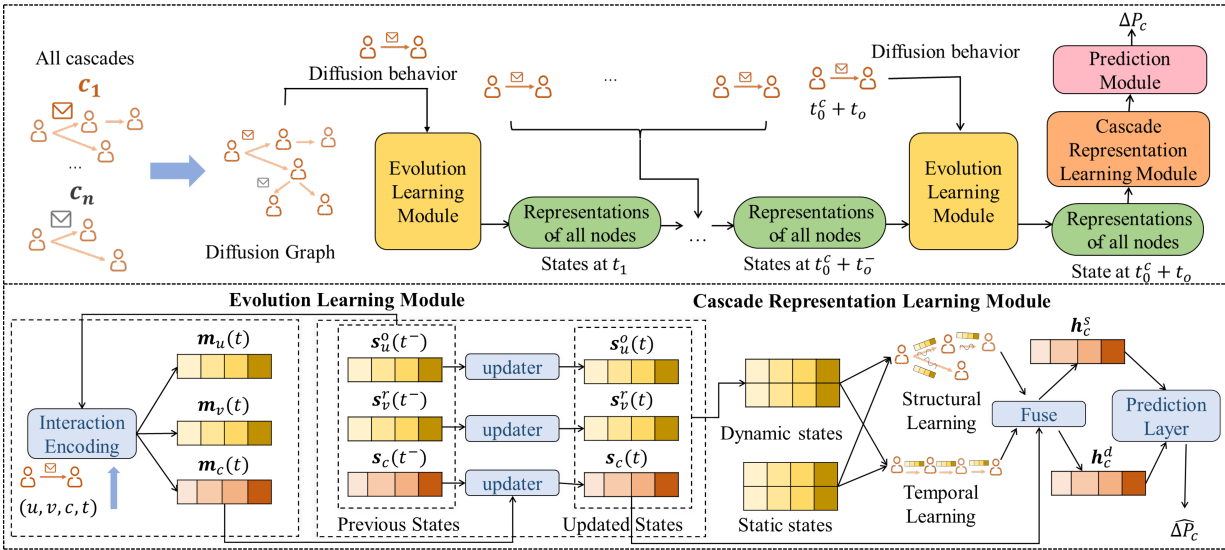


Figure 2: Framework of the proposed method. It consists of an evolution learning module to model the dynamics of user preferences and the correlation between cascades, a cascade representation learning module to capture the temporal and structural information within a cascade, and a prediction module to give future popularity. The popularity of cascade is predicted according to the most recent representations.

$(u_1, t_1), (u_2, t_2), \dots, (u_n, t_n)$ where (u_i, t_i) indicates that user u_i participate in the cascade c at t_i after the publication of the cascade. The target of this module is to learn the temporal pattern from the diffusion sequence such as the short-term outbreak of user participation. The direct way to learn the temporal pattern is feeding participating users' representations sequentially into a recurrent neural network, however, it may neglect the time information in the diffusion sequence since it can not distinguish users participating at different times. Inspired by the position embedding technics [Vaswani *et al.*, 2017], we divide the observation time t_o into n_t slots $[0, \frac{t_o}{n_t}), [\frac{t_o}{n_t}, 2\frac{t_o}{n_t}), \dots, [(n_t - 1)\frac{t_o}{n_t}, t_o)$ and preserve a learnable embedding e_i^t for every time interval $[\frac{t_o}{n_t}, (i + 1)\frac{t_o}{n_t})$ to distinguish users that participate in the cascade at different time. Besides, we also introduce another learnable parameter e^p to strengthen the path information, where e_i^p is a position embedding for the i th participating users. We get the user embedding $z_{u_i}^s$ by first adding these two embeddings to the states of users and then feeding the sequence of user embeddings to the Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] to get the cascade temporal embedding $h_{c,t}^s$, that is,

$$h_{c,t}^s = LSTM^s([z_{u_1}^s, z_{u_2}^s, \dots, z_{u_n}^s]), \quad (4)$$

$$z_{u_i}^s = s_{u_i} + e_{[t_i]}^t + e_i^p, \quad (5)$$

where $[t_i]$ is the time slot that t_i belongs to, i.e., $[t_i] * \frac{t_o}{n_t} \leq t_i < ([t_i] + 1) * \frac{t_o}{n_t}$. Here the superscript of $h_{c,t}^s$ means it is the static temporal representation and we also generate a dynamic temporal representation $h_{c,t}^d$ by the above equation except using different LSTM parameters and user representations (i.e., dynamic states).

Structural Learning. Besides the order and time interval of participating users, the cascade graph also plays an important role in popularity prediction. For example, a deeper

cascade may get more popularity since it influences the users who are far away from the original user [Cheng *et al.*, 2014]. The cascade graph can be considered as a directed acyclic graph (DAG) with a root node (the user who posts the message), where a path from the root node to other nodes represents a diffusion process of a message in the social network. Though graph neural networks like GCN can learn graph structure, it may be difficult to model deep cascade paths [Tang *et al.*, 2021]. Inspired by [Tai *et al.*, 2015; Ducci *et al.*, 2020], we employ a modified LSTM and aggregate representations of users on the cascade graph along the direction of information flow. Formally, let $S(u)$ and $T(u)$ be the users that u receives messages from and sends messages to, i.e., there are edges pointing from $S(u)$ to u and u to $T(u)$. Then we employ the following mechanism to propagate the information from root nodes to leaf nodes.

$$\begin{aligned} \tilde{h}_{u,\uparrow}^s &= \sum_{v \in S(u)} h_{v,\uparrow}^s, \\ \mathbf{i}_{u,\uparrow}^s &= \sigma(\mathbf{W}_{i,\uparrow}^s [s_u || \tilde{h}_{u,\uparrow}^s] + \mathbf{b}_{i,\uparrow}^s), \\ \mathbf{f}_{uv,\uparrow}^s &= \sigma(\mathbf{W}_{f,\uparrow}^s [s_u || h_{v,\uparrow}^s] + \mathbf{b}_{f,\uparrow}^s), \\ \mathbf{o}_{u,\uparrow}^s &= \sigma(\mathbf{W}_o^s [s_u || \tilde{h}_{u,\uparrow}^s] + \mathbf{b}_{o,\uparrow}^s), \\ \mathbf{g}_{u,\uparrow}^s &= \tanh(\mathbf{W}_{g,\uparrow}^s [s_u || \tilde{h}_{u,\uparrow}^s] + \mathbf{b}_{g,\uparrow}^s), \\ \mathbf{c}_{u,\uparrow}^s &= \mathbf{i}_{u,\uparrow}^s \odot \mathbf{g}_{u,\uparrow}^s + \sum_{v \in S(u)} \mathbf{f}_{uv,\uparrow}^s \odot \mathbf{c}_{v,\uparrow}^s, \\ \mathbf{h}_{u,\uparrow}^s &= \mathbf{o}_{u,\uparrow}^s \odot \tanh(\mathbf{c}_{u,\uparrow}^s), \end{aligned} \quad (6)$$

After propagating the information in the graph, we sum the leaf nodes' representations to get the cascade embedding $h_{c,\uparrow}^s = \sum h_{leaf,\uparrow}^s$. Besides, we reverse the edge direction of the cascade graph and generate another cascade representation $h_{c,\downarrow}^s$ from leaf to root. Finally, we concatenate the $h_{c,\uparrow}^s$ and $h_{c,\downarrow}^s$ and feed it to an MLP to get the final structural rep-

resentation $\mathbf{h}_{c,s}^s$. Here the superscript of $\mathbf{h}_{c,s}^s$ represents the static structural embedding of the cascade as in the temporal learning module. We also generate the dynamic representation $\mathbf{h}_{c,s}^d$ by the same mechanism in Equation (6) except using different parameters and user representations.

Embedding Fusion. In this module, we fuse the temporal embedding and structural embedding into a cascade embedding. For static embedding $\mathbf{h}_{c,t}^s$ and $\mathbf{h}_{c,s}^s$, we get the merged embedding \mathbf{h}_c^s by concatenating them and then feed it into an MLP. The merge process of the dynamic embedding is slightly different from that of the static, where we split the participating users into two parts: the users u and v participating in the last diffusion (u, v, c, t) of a cascade c and others. The last two users u, v 's dynamic states are used to merge with the dynamic cascade state $\mathbf{s}_c(t)$ and the others are used to generate the temporal and structural embedding $\mathbf{h}_{c,t}^d$ and $\mathbf{h}_{c,s}^d$. The reason for this is that the last two users' dynamic states are updated from $\mathbf{s}_u^o(t^-), \mathbf{s}_v^r(t^-)$ to $\mathbf{s}_u^o(t), \mathbf{s}_v^r(t)$ by the updaters in (3) and this makes the gradients can be propagated back to the updaters through them, which makes them different from the dynamic states of other users. Formally, the merge process of the dynamic representation can be represented as

$$\mathbf{h}_c^d = \sigma(\mathbf{W}_a[\mathbf{h}_{c,t}^d \parallel \mathbf{h}_{c,s}^d \parallel \tilde{\mathbf{h}}_c^d]), \quad (7)$$

$$\tilde{\mathbf{h}}_c^d = \sigma(\mathbf{W}_b[\tilde{\mathbf{s}}_c(t) \parallel \mathbf{s}_u^o(t) \parallel \mathbf{s}_v^r(t)]), \quad (8)$$

$$\tilde{\mathbf{s}}_c(t) = \mathbf{s}_c(t) + \mathbf{e}_{[t_0^c]}^g, \quad (9)$$

where t_0^c is the publication time of c and \mathbf{e}^g is another position embedding for publication time like Equation (5).

4.3 Prediction Module

In this module, we give the prediction of incremental popularity by merging the prediction result from static embedding \mathbf{h}_c^s and dynamic embedding \mathbf{h}_c^d .

$$\widehat{\Delta P}_c = \lambda f_{\text{static}}(\mathbf{h}_c^s) + (1 - \lambda) f_{\text{dynamic}}(\mathbf{h}_c^d), \quad (10)$$

where the $f_{\text{static}}(\cdot)$ and $f_{\text{dynamic}}(\cdot)$ are two MLP functions and λ is a hyperparameter to control the weight of static result and dynamic result.

We use the Mean Squared Logarithmic Error (MSLE) as the loss function, which can be formulated as follows,

$$\mathcal{J}(\theta) = \frac{1}{n} \sum_c (\log(\Delta P_c) - \log(\widehat{\Delta P}_c))^2, \quad (11)$$

where n is the number of training cascades.

5 Experiments

In this section, we conduct experiments on three datasets to evaluate the effectiveness of our approach.

5.1 Descriptions of Datasets

We use three real-world datasets in the experiments, including the cascades in social platforms (Twitter and Weibo) and academic networks (APS).

- **Twitter** [Weng *et al.*, 2013] contains the tweets published between Mar 24 and Apr 25, 2012 on Twitter and their retweets during this period. Every cascade in this dataset represents the diffusion process of a hashtag.

- **Weibo** [Cao *et al.*, 2017] was collected on Sina Weibo which is one of the most popular Chinese microblog platform. It contains posts published on July 1st, 2016 and their retweets during this period. Every cascade in this dataset represents the diffusion process of a post.
- **APS**¹ contains papers published on American Physical Society (APS) journals and their citation relationships before 2017. Every cascade in this dataset represents the process of obtaining citations for a paper. Following previous works [Cao *et al.*, 2017], transformation and preprocessing are taken to make paper citation prediction analogy to the retweet prediction.

Following Xu *et al.* [2021], we randomly select 70%, 15% and 15% of the cascades for training, validating and testing. For data preprocessing, we set the observation window of a cascade to 2 days, 1 hour and 5 years on Twitter, Weibo and APS. For Weibo and Twitter, we predict cascades' popularity at the end of the dataset, while we predict cascades' popularity 20 years after its publication for APS. The cascades whose observed popularity $|c(t_0^c + t_o)|$ is less than 10 are discarded and for cascades whose $|c(t_0^c + t_o)|$ is more than 100, we only select the first 100 participants. Moreover, to ensure that there are adequate times for cascades to accumulate popularity and to avoid the effect of diurnal rhythm [Cao *et al.*, 2017], we select the cascades published before April 4th, published between 8:00 and 18:00, and published before 1997 on Twitter, Weibo and APS, respectively. The above preprocessing process also follows previous methods [Xu *et al.*, 2021; Cao *et al.*, 2017]. Table 1 shows the statistics of the datasets.

Datasets	#Users	#Cascades	#Retweets
Twitter	199,005	19,718	602,253
Weibo	918,852	39,076	1,572,287
APS	218,323	48,575	939,686

Table 1: Statistics of datasets.

5.2 Baselines

We compare our method with the following baselines, where the first two methods (i.e., XGBoost and MLP) additionally need hand-designed features (see details in Section 5.4):

- **XGBoost** belongs to the gradient boosting algorithm, which is a widely used machine learning method [Chen and Guestrin, 2016].
- **MLP** uses the multilayer perceptron to compute on the features of each cascade.
- **DeepHawkes** [Cao *et al.*, 2017] treats each cascade as multiple diffusion paths of users and learns sequential information of cascades through the GRU.
- **DFTC** [Liao *et al.*, 2019] considers each cascade as a popularity count sequence and uses the Convolutional Neural Network (CNN), LSTM and attention mechanism to learn the cascade representation.

¹<https://journals.aps.org/datasets>

- **MS-HGAT** [Sun *et al.*, 2022] builds a sequence of regularly-sampled hypergraphs that contain multiple cascades and users, and then learns on hypergraphs for computing the representations of cascades.
- **CasCN** [Chen *et al.*, 2019b] treats each cascade as a graph sequence and uses the GNN and LSTM to learn cascade representations.
- **TempCas** [Tang *et al.*, 2021] additionally designs a sequence modeling method to capture macroscopic temporal patterns apart from learning on the cascade graph.
- **CasFlow** [Xu *et al.*, 2021] is the state-of-the-art method for cascade prediction, which first learns users’ representations from the social network and the cascade graph and then employs the GRU and Variational AutoEncoder (VAE) to get representations of cascades.

5.3 Evaluation Metrics

We choose four widely used metrics to evaluate the performance of the compared methods, including Mean Squared Logarithmic Error (MSLE), Mean Absolute Logarithmic Error (MALE), Mean Absolute Percentage Error (MAPE) and Pearson Correlation Coefficient (PCC). Among these metrics, MSLE, MAPE and MALE evaluate the prediction error between the predicted value and the ground truth from different aspects and PCC measures the correlation between predicted value and the ground truth.

5.4 Experimental Settings

For XGBoost and MLP, we follow Cheng *et al.*[2014] and extract five types of features (i.e., edge number, max depth, average depth, breath of cascade graph, and publication time of the cascade) as the hand-designed cascade features. We set the dimension of dynamic states of users and cascades, as well as the cascade embedding to 64. The dimension of position embedding is set to 16. The time slot number n_t is set to 20 and the fusion weight λ is 0.1. For training, we adopt the Adam optimizer and use the early stopping strategy with a patience of 15. The learning rate and batch size are set to 0.0001 and 50. Our code can be found at <https://github.com/lxd99/CTCP>.

5.5 Performance Comparison

Table 2 reports the performance of different methods, and some conclusions can be summarized as follows.

Among the three groups of methods, feature-based models perform the worst among all baselines, which reveals that there are complex evolution patterns of the cascade size that can not be captured by the hand-designed features. Moreover, graph-based models show better performance than sequence-based models, implying the necessity of exploiting the structural and temporal information carried in the cascade graph.

CTCP achieves significant performance improvement w.r.t. the state-of-the-art baseline (i.e., CasFlow) on Twitter and APS, demonstrating the effectiveness of the proposed method. This improvement may be due to the fact that we learn the dynamic representations of cascades and users collaboratively, which can capture the correlation between cascades and the dynamic user preferences outside of a single

cascade. The insignificant improvement of CTCP on Weibo may be due to the short time period of Weibo (1 day compared to 1 month and more than 100 years on Twitter and APS respectively) and the preferences of users may not evolve during such a short period, which makes CTCP have no advantages over CasFlow. Additionally, modeling multiple cascades via the sequence of graph snapshots like MS-HGAT does not achieve considerable performance. Because the diffusion behaviors within a snapshot are thought to happen at the same time which will lose fine-grained temporal information. Moreover, MS-HGAT needs to load the snapshot into memory at one time, which makes it can only run on the smallest dataset (i.e., Twitter).

5.6 Sensitivity to Publication Time

To explore the sensitivity of different models to the publication time of cascades, we plot models’ performance on cascades with different publication times on Twitter and APS. Specifically, we divide the cascade into five groups according to their publication time: cascade whose publication time is at the 0th to 20th, 20th to 40th, 40th to 60th, 60th to 80th and 80th to 100th percentile, and plot the best five models’ performance. From Figure 3, we can observe that CTCP can achieve considerable performance on different cascades consistently. Besides, as time goes on, the performance of CTCP consistently improves on these two datasets. This is because the evolution learning module of CTCP keeps updating the dynamic states of users and as time goes on more and more user behaviors are observed, which provides richer information to model the preference of users. Other models only learn from the own diffusion process of cascades and can not learn this dependency.

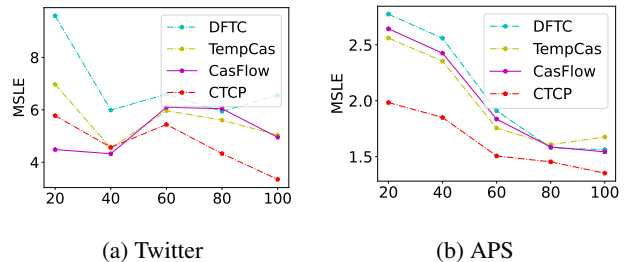


Figure 3: Performance on cascades with different publication times. The horizontal axis indicates the percentile range of the publication time of cascades and the vertical axis is the average prediction MSLE of cascade in that range.

5.7 Ablation Study

We compare CTCP with the following variations on Twitter and APS to investigate the contribution of submodules to the prediction performance.

- **w/o EL** removes the evolution learning module.
- **w/o SE** removes the static representation of users.
- **w/o SL**: removes the structural learning module in the cascade embedding learning process.

Model	Twitter				Weibo				APS			
	MSLE	MAE	MAPE	PCC	MSLE	MAE	MAPE	PCC	MSLE	MAE	MAPE	PCC
XGBoost	11.5330	2.9871	0.8571	0.3792	3.6253	1.3736	0.3571	0.6493	2.5808	1.2559	0.3437	0.4762
MLP	11.9105	2.9712	0.9324	0.3733	3.9370	1.4409	0.3812	0.6098	2.6075	1.2577	0.3516	0.4787
DeepHawkes	7.7795	2.1553	0.5547	0.6500	4.2520	1.4658	0.3998	0.5670	2.3356	1.2001	0.3158	0.5524
DFTC	5.9173	1.8426	0.4851	0.7495	2.9370	1.2046	0.2959	0.7296	2.0357	1.1159	0.2943	0.6247
CasCN	7.1021	2.0567	0.5231	0.6940	3.7714	1.4040	0.3612	0.6707	2.1248	1.1358	0.3035	0.6062
MS-HGAT	5.9992	1.9006	0.4741	0.7507	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
TempCas	5.5870	1.7584	0.4574	0.7651	2.7453	1.1702	0.2786	0.7500	2.0043	1.1022	0.2957	0.6346
CasFlow	5.2549	1.5775	0.4031	0.7847	2.6336	1.1230	0.2687	0.7619	2.0064	1.1053	0.2936	0.6320
CTCP	4.6916	1.5668	0.3562	0.8136	2.5929	1.1414	0.2723	0.7667	1.6289	0.9906	0.2611	0.7176

Table 2: Performance of all methods in three datasets, where the methods can be divided into three categories: feature-based, sequence-based, and graph-based methods from top to bottom in the table. The best results appear in bold and OOM indicates the out-of-memory error.

From Figure 4, we can observe that: Firstly the performance of w/o EL and w/o SE varies on APS and Twitter, for example, w/o SE achieves the best performance on Twitter and the worst performance on APS. This indicates that the growth of the cascade size is controlled by multiple factors and it is necessary to consider the dynamic preference and static preference of users simultaneously. Secondly, the structural learning module utilizes the cascade graph to generate the cascade embedding which helps improve the prediction performance by capturing the evolution pattern of a cascade at a macro level.

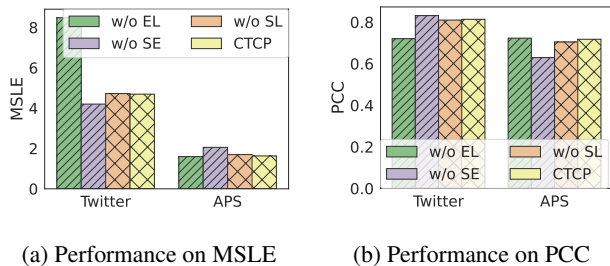


Figure 4: Ablation study on Twitter and APS.

5.8 Cascade Representations Projection

To confirm the effectiveness of the learned cascade representations, we project the cascade representations of CTCP and CasFlow on Twitter into a two-dimensional space, using t-NSE [van der Maaten and Hinton, 2008]. Results are represented in Figure 5. Remarkably, we find that the learned representations of CTCP can capture the evolution pattern of cascade popularity, suggested by the fact that from right-top to left-bottom the node color of CTCP changes from red to dark blue continuously in Figure 5 (a). While for CasFlow, nodes with different colors are mixed. This may be because CTCP models the correlation of cascades while CasFlow does not, which can help the model capture the collaborative signals between cascades and learn a better cascade representation.

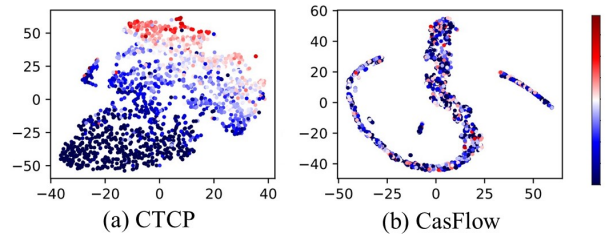


Figure 5: Projection of learned cascade representations on Twitter, where each point represents a cascade representation and the color represents its incremental popularity. (Dark blue means low popularity and dark red means high popularity).

6 Conclusion

In this paper, we studied the problem of cascade popularity prediction and pointed out two factors that are not considered well in the existing methods, i.e., the correlation between cascades and the dynamic preferences of users. Different from previous methods that independently learn from each cascade, our method first combines all cascades into a diffusion graph to explore the correlations between cascades. To model the dynamic preferences of users, an evolution learning module was proposed to learn on the diffusion graph chronologically, which maintains dynamic states for users and cascades, and the states are updated continuously once a diffusion behavior happens. Moreover, a cascade representation learning module was proposed to explore the structural and temporal information within a cascade by aggregating representations of users into a cascade embedding. Extensive experimental results on three real-world datasets demonstrated the effectiveness of the proposed method.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments on this research. This work was supported by the National Key R&D Program of China (2021YFB2104802) and the National Natural Science Foundation of China (62272023).

References

- [Cao *et al.*, 2017] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1149–1158. ACM, 2017.
- [Chang *et al.*, 2020] Xiaofu Chang, Xuqin Liu, Jianfeng Wen, Shuang Li, Yanming Fang, Le Song, and Yuan Qi. Continuous-time dynamic graph learning via neural interaction processes. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 145–154. ACM, 2020.
- [Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.
- [Chen *et al.*, 2019a] Xueqin Chen, Kunpeng Zhang, Fan Zhou, Goce Trajcevski, Ting Zhong, and Fengli Zhang. Information cascades modeling via deep multi-task learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 885–888. ACM, 2019.
- [Chen *et al.*, 2019b] Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. Information diffusion prediction via recurrent cascades convolution. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 770–781. IEEE, 2019.
- [Cheng *et al.*, 2014] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 925–936. ACM, 2014.
- [Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014.
- [Ducci *et al.*, 2020] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel. Cascade-1stm: A tree-structured neural classifier for detecting misinformation cascades. In Rakesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2666–2676. ACM, 2020.
- [Guo and Suo, 2014] Jin-Li Guo and Qi Suo. Comment on "quantifying long-term scientific impact". *CoRR*, abs/1405.1574, 2014.
- [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2704–2710, 2020.
- [Huang *et al.*, 2021] Han Huang, Leilei Sun, Bowen Du, Chuanren Liu, Weifeng Lv, and Hui Xiong. Representation learning on knowledge graphs for node importance estimation. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 646–655. ACM, 2021.
- [Kazemi *et al.*, 2020] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs: A survey. *J. Mach. Learn. Res.*, 21:70:1–70:73, 2020.
- [Leskovec *et al.*, 2007] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.
- [Li *et al.*, 2017] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 577–586. ACM, 2017.
- [Liao *et al.*, 2019] Dongliang Liao, Jin Xu, Gongfu Li, Weijie Huang, Weiqing Liu, and Jing Li. Popularity prediction on online articles with deep fusion of temporal process and content features. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 200–207. AAAI Press, 2019.
- [Pareja *et al.*, 2020] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao B. Schardl, and Charles E. Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5363–5370. AAAI Press, 2020.
- [Rossi *et al.*, 2020] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. *CoRR*, abs/2006.10637, 2020.
- [Sankar *et al.*, 2020] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. Dysat: Deep neural representation learning on dynamic graphs via self-attention

- networks. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 519–527. ACM, 2020.
- [Schlichtkrull *et al.*, 2018] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, 2018.
- [Sun *et al.*, 2022] Ling Sun, Yuan Rao, Xiangbo Zhang, Yuqian Lan, and Shuanghe Yu. MS-HGAT: memory-enhanced sequential hypergraph attention network for information diffusion prediction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 4156–4164. AAAI Press, 2022.
- [Szabó and Huberman, 2010] Gábor Szabó and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, 2010.
- [Tai *et al.*, 2015] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566. The Association for Computer Linguistics, 2015.
- [Tang *et al.*, 2021] Xiangyun Tang, Dongliang Liao, Weijie Huang, Jin Xu, Liehuang Zhu, and Meng Shen. Fully exploiting cascade graphs for real-time forwarding prediction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 582–590. AAAI Press, 2021.
- [van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [Wang *et al.*, 2019] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2022–2032, 2019.
- [Wang *et al.*, 2020] Xiao Wang, Deyu Bo, Chuan Shi, Shaohua Fan, Yanfang Ye, and Philip S. Yu. A survey on heterogeneous graph embedding: Methods, techniques, applications and sources. *CoRR*, abs/2011.14867, 2020.
- [Wang *et al.*, 2021] Ruijie Wang, Zijie Huang, Shengzhong Liu, Huajie Shao, Dongxin Liu, Jinyang Li, Tianshi Wang, Dachun Sun, Shuochao Yao, and Tarek F. Abdelzaher. Dydiff-vae: A dynamic variational framework for information diffusion prediction. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 163–172. ACM, 2021.
- [Weng *et al.*, 2013] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6, 2013.
- [Wu *et al.*, 2019] Qitian Wu, Yirui Gao, Xiaofeng Gao, Paul Weng, and Guihai Chen. Dual sequential prediction models linking sequential recommendation and information dissemination. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 447–457. ACM, 2019.
- [Xu *et al.*, 2020] Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [Xu *et al.*, 2021] Xovee Xu, Fan Zhou, Kunpeng Zhang, Siyuan Liu, and Goce Trajcevski. Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.
- [Yang *et al.*, 2019] Cheng Yang, Jian Tang, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. Multi-scale information diffusion prediction with reinforced recurrent networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4033–4039. ijcai.org, 2019.