DAG-Aware Variational Autoencoder for Social Propagation Graph Generation

Dongpeng Hou[1, 2], **Chao Gao**[2*], **Xuelong Li**[2], **Zhen Wang**[3, 1, 2†]

[1]School of Mechanical Engineering, Northwestern Polytechnical University
[2]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
[3]School of Cybersecurity, Northwestern Polytechnical University

Abstract

Propagation models in social networks are critical, with extensive applications across various fields and downstream tasks. However, existing propagation models are often oversimplified, scenario-specific, and lack real-world user social attributes. These limitations detaching from real-world analysis lead to inaccurate representations of the propagation process in social networks. To address these issues, we propose a User Features Attention-based **DAG**-**A**ware **V**ariational **A**utoencoder (DAVA) for propagation graph generation. First, nearly 1 million pieces of user attributes data are collected. Then DAVA can integrate the analysis of propagation graph topology and corresponding user attributes as prior knowledge. By leveraging a lightweight attention-based framework and a sliding window mechanism based on BFS permutations weighted by user influence, DAVA significantly enhances the ability to generate realistic, large-scale propagation data, yielding graph scales ten times greater than those produced by existing SOTA methods. Every module of DAVA has flexibility and extension that allows for easy substitution to suit other generation tasks. Additionally, we provide a comprehensive evaluation of DAVA, one focus is the effectiveness of generated data in improving the performance of downstream tasks. During the generation process, we discover the Credibility Erosion Effect by modifying the generation rules, revealing a social phenomenon in social network propagation.

Introduction

In recent years, the analysis of propagation models in social networks has attracted growing attention due to their considerable impact on various aspects of society (Leskovec et al. 2007; Vosoughi, Roy, and Aral 2018). Numerous applications within the realm of social networks are fundamentally based on the propagation process. Examples include influence assessment (Xia et al. 2021), locating the diffusion source (Wang et al. 2023), and user profiling (Jiang, Ren, and Ferrara 2023), all of which rely on different propagation models. Consequently, by employing these models in conjunction with downstream tasks, far-reaching implications can be observed across a wide range of fields, such as politics, public health, and marketing (Goel et al. 2016).

*Corresponding author: cgao@nwpu.edu.cn
†Corresponding author: w-zhen@nwpu.edu.cn

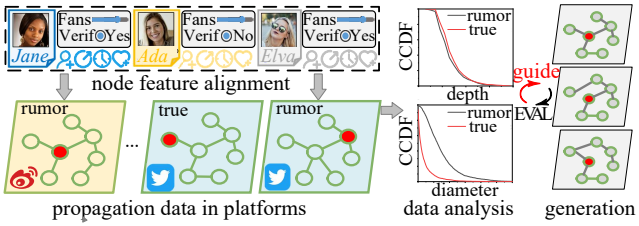

Figure 1: The illustration of a novel propagation graph generation grounded in real-world propagation data. Recognizing that propagation is user-driven, we analyze the propagation data in social media, emphasizing unique user attributes and propagation structures. These analyses guide our graph generation closely mirroring real-world propagation.

One of the widely used traditional propagation models in the social network is information cascade models, which provide a basis for understanding and analyzing the spread of information, and behaviors in various scenarios (Shakarian et al. 2015). However, these propagation models are often tailored for specific scenarios and depend on simplistic assumptions. Further, some deep learning submodules, easily integrated into downstream tasks, adeptly capture complex relationships and patterns in information propagation (Xia et al. 2021; Ling et al. 2022). However, propagation models based on representation learning in deep frameworks are often latent and lack direct interpretability, making it difficult to understand the underlying mechanisms.

In summary, the most significant drawback of the aforementioned methods is their detachment from real-world data (Chen, Castillo, and Lakshmanan 2022; Guille et al. 2013), which leads to a weak capacity to characterize propagation dynamics in actual scenarios. Therefore, the motivation of this paper is to conduct research based on real propagation data from social media and propose a framework that effectively reflects the real propagation process. However, there are two existing challenges. First, we recognize that propagation in social networks is user-driven (Li et al. 2022), but existing propagation data barely contains user information, resulting in propagation based directed acyclic graph (DAG) in social media that only has a topological structure without assigning user attributes to nodes. Second, the cost of obtaining real propagation datasets is high, and due to

hardware limitations, the scale of the collected propagation data is generally small (such as the user scale in the Twitter propagation dataset is only smaller than 2,000) (Liu et al. 2015; Ma et al. 2016), which is not conducive to demonstrating the performance of downstream algorithms on different scale datasets.

In order to overcome these two challenges, we have made some efforts. First, we crawl nearly one million user attributes, including the number of tweets, followers, and followings, then correlate these attributes to the nodes in the real propagation data based on the unique UID reliably, ensuring the completeness and authenticity of propagation events. Then a comprehensive and thorough analysis of the Twitter and Weibo propagation data, including various structural topologies and user features, is conducted to obtain the propagation characteristics distribution. Second, based on these prior information, we further introduce a graph generative model to learn the propagation patterns of both rumors and non-rumors[1]. More specifically, a User Features Attention based DAG-Aware Variational Autoencoder (DAVA) is developed for the propagation graph generation in the social network. DAVA employs an exponential distribution sampling based variational autoencoder for graph-level generation and a relationship attention mechanism focused on user attributes for edge-level generation. Finally, DAVA can construct large-scale propagation datasets with similar characteristics as those statistically observed on Weibo and Twitter. The contributions of this paper are summarized below.

- We provide a comprehensive analysis of real-world propagation data from Twitter and Weibo platforms with nearly one million crawled users. Incorporating these insights directly DAVA can improve the quality of the generated output and reduce the depth of hidden layer. And the code is available at https://github.com/cgao-comp/DAVA.
- We introduce an innovative model DAVA for social propagation data generation. It features a graph permutation with user importance, exponential sampling in the graph autoencoder, an interpretable attention mechanism that focuses on user relationships, and a unique loss function. What's more, the time-sliding window strategy generates graphs ten times larger than SOTA methods.
- We identify a phenomenon called Credibility Erosion Effect in social network propagation. Importantly, we incorporate this discovery into DAVA's generation process by applying a decay factor to the predecessors. Such a factor mirrors the credibility erosion effect, enhancing the realism and effectiveness of the generative graphs.
- We expand the evaluation strategies for DAVA to rigorously verify the generation ability similar to real-world propagation characteristics, including traditional metrics in the generative field, comparing feature distributions with real data using CCDF, and assessing utility in downstream tasks like source localization.

[1]We adopt the traditional definition from the social psychology literature (Allport and Postman 1947), which defines a rumor as a story or statement whose truth value is unverified or deliberately false, for better understanding.

Related Work

In this paper, we employed deep graph generative techniques to generate and augment real-world propagation datasets in social networks. Therefore, the related work should be discussed in two parts: propagation models in social networks and deep graph generative models.

Propagation Models

The propagation models are helpful to comprehend how sources are formed, information is spread, and group behaviors are influenced in social networks. Some classical influence diffusion models are widely used to characterize the propagation process in social networks. For example, Kempe et al. propose the Independent Cascade (IC) and Linear Threshold (LT) models (Kempe, Kleinberg, and Tardos 2003). Following the successful application of the Susceptible-Infectious (SI) and Susceptible-Infectious-Recovered (SIR) models in epidemiology, these models have been adopted in social networks (Wang et al. 2022). However, these models often make simplified assumptions and are restricted to specific scenarios, limiting their real-world applicability. To completely learn the complex interaction parameters of the underlying propagation models, methods based on deep learning have received widespread attention in recent years. The IVGD model is a versatile architecture of reversible graph diffusion models, designed to autonomously learn the inherent rules of the propagation process (Wang, Jiang, and Zhao 2022). Despite deep methods having an adept ability to learn heterogeneous parameters of propagation models, their robustness and transferability are challenged without the support of real-world propagation analysis (Wu et al. 2020).

Deep Graph Generative Models

Deep graph generative techniques are initially successfully applied in fields like chemistry and pharmaceuticals (You et al. 2018a; Jin, Barzilay, and Jaakkola 2018). For instance, Li et al. use Graph Convolutional Networks (GCN) to capture both the structure and attributes during the generation process (Li et al. 2018). With these mature applications in various domains, such techniques have been extended to social networks. GraphRNN, for instance, address the limitations of previous methods that could only learn from a single graph or generate small-scale graphs (You et al. 2018b). However, GraphRNN primarily focuses on graph permutation embedding and the scale of generation, rather than adequately learning the attribute information within the graph. Zhang et al. further consider the attribute-based neural architecture graphs and introduced the DAG based variational autoencoder D-VAE (Zhang et al. 2019). This approach allowed for more comprehensive learning and generation of attribute information within graphs. And the goal of D-VAE bears resemblance to our task of generating social network propagation DAG. However, D-VAE falls short in adequately distinguishing the weight relationships between nodes by employing a unified aggregation based on message passing. Additionally, current works (Han et al. 2023; Zahirnia et al. 2022) entail relatively high computational complexity so it is difficult to generate a large-scale graph.

Methodology

Problem Definition

Considering a DAG $G = (V, E, F)$, which represents a propagation process related to a unique event topic extracted via web crawling from social platforms like Twitter or Weibo. Here, $V = \{v_1, ..., v_n\}$ represents the set of users, $E = \{(v_i, v_j) \mid i \neq j, (v_i, v_j) \in V \times V\}$ symbolizes the propagation paths, where each edge from v_i to v_j signifies that a piece of information is disseminated from user v_i to user v_j, $F = \{f(v) \mid v \in V\}$ denotes the feature vectors associated with each user, such as user profile information or posting behaviors. The goal is to construct a generative model M that can accurately mimic the statistical properties of the extracted propagation network G. Specifically, the model should be capable of generating a new network $G' = (V', E', F')$ where V', E', and F' are the sets of users, propagation paths, and user features in the generated network, respectively, The generated network G' should exhibit similar statistical properties on some rigorous metrics to the extracted network G.

User Features Attention Based DAG-Aware Variational Autoencoder

Fast Preprocessing for Propagation DAG Different types (rumor and non-rumor) of propagation DAGs exhibit significant disparities in both features and structures. More specifically, leveraging social data analysis techniques, we used the Complementary Cumulative Distribution Function (CCDF) for evaluation, examining Weibo and Twitter's propagation data across aspects like network diameter, propagation depth and breadth, and structural virality. Our findings demonstrate that rumors spread more virally, whereas non-rumors disperse in a broadcast-like fashion. An analysis of user attributes also showed that source users broadcasting non-rumors are typically of higher quality, while those spreading rumors are more active[2]. And these insights guide our proposed generative model. One focus is an abundance of users in the propagation DAGs that are directly linked to the propagation source and have an out-degree of zero. The impact of these users on the propagation structure is relatively minor given that their depth is fixed at 2, thereby contributing weakly to the structural virality. We utilize a simple yet effective model $f_\theta(G)$ to learn the number of such users, denoted as $\hat{V}$. By eliminating these connections prior to the graph generation training process, we can decrease redundant information, streamline the dataset, and improve training efficiency.

$$f_\theta(G(V - \hat{V}, E - \hat{E}, F)) \rightarrow |\hat{V}|, \tag{1}$$

where $\forall v \in \hat{V}$, $out\text{-}neighbor(v)$=0 and $(s, v) \in E$. s is the source user in a DAG. Then the generative model could not focus on these meaningless users and their related connections, and the costs arising from these details can be directly omitted. Specifically, we leverage the graph pooling technique, Set2Set (Vinyals, Bengio, and Kudlur 2015), to directly encode the entire DAG, including its topology and node features, thereby implementing Eq. (1).

$$\begin{aligned} f_\theta(G) &\triangleq f_\theta(G(V - \hat{V}, E - \hat{E}, F)) \\ &= \text{Nonlinear}(\text{Set2Set}(G(V - \hat{V}, E - \hat{E}, F))). \end{aligned} \tag{2}$$

Note that the choice of the function f_θ is flexible. Some other models with aggregation capabilities, such as GCN (Welling and Kipf 2016), GAT (Velickovic et al. 2018), etc., can also be applied to f_θ. It's worth mentioning that when finally generating a new graph G', we only need to perform $f_\theta(G')$ and add these users and connections.

One of the biggest challenges in graph generation is non-unique representations. A graph with n nodes can correspond to up to n equivalent adjacency matrices due to arbitrary node orderings, making it computationally expensive to model and optimize objective functions of graph generation. Therefore, based on user features E, we propose a unique graph representation method based on node importance using Breadth-First Search (BFS) to ensure a unique node index sequence Φ for the matrix representation of identical graph structures. Due to space constraints, we demonstrate the implementation process in Alg. 1. For more detailed reference, see the publicly available source code.

Algorithm 1: BFS Permutation with Node Importance

Require: A propagation DAG $G(V, E, F)$
Ensure: The BFS sequence $bfsSequence$ (i.e., Φ)

$F = MinMaxScaler(F)$ // Normalization
$Chi\text{-}square(F) \rightarrow \{f'_1, f'_2, ...\}$ // Sort the importance of different features.
Provide a unique one-hot encoding representation $I(v)$ for each user based on the sorted dimension$\{f'_1, f'_2, ...\}$.
Initialize $visited \leftarrow \emptyset$, $curLevel \leftarrow [root]$, $bfsSequence \leftarrow []$
while $curLevel \neq \emptyset$ **do**
  Sort $curLevel$ based on feature importance $I(v)$
  $nextLevel \leftarrow \emptyset$
  for $v_i \in curLevel$ **do**
    if $v_i \notin visited$ **then**
      $visited$.add(v_i), $bfsSequence$.add(v_i)
      $nextLevel$.add($Neighbor\_list(v_i)$)
    end if
  end for
  $curLevel \leftarrow nextLevel$, $nextLevel \leftarrow \emptyset$
end while
return $bfsSequence$

[2]Our comprehensive analysis reveals some phenomena divergent from previous research. For instance, rumor spread is slower but persists longer; 90% rumors involve fewer participants than non-rumor events. However, the remaining rumors, though rare, are quite sensational. These rumors attract a significant number of participants, leading to a higher average than median participant count from an overall perspective. Additionally, we identify that reputable active users, termed 'onlookers', inadvertently or unwittingly spread rumors due to their extensive online interactions and the allure of sensational fake news. Conversely, celebrities exhibit caution, mindful of releasing unverified information. More detailed information can be found in an analysis (Hou et al. 2024).

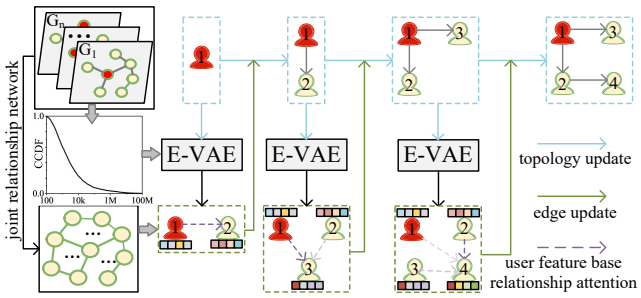

Figure 2: The illustration of the generation process in DAVA. The graph-level generation process (blue line) is primarily centered on the topological structure of DAGs. This process employs an exponential variational autoencoder to establish latent space vector representations of the graph structures. Building upon this foundational topology, the edge-level generation process (green line) evaluates the similarity between a potential new user and each existing user. For such an evaluation, DAVA leverages a user attribute attention mechanism (purple line) to assess the likelihood of directed edge formation between the new user and each existing user. Then the newly generated edge structures are subsequently integrated into the next sequence of graph-level generation.

The Generation Process So far, the unique input sequence Φ can be determined for the same graph. Next, we would like to introduce the graph generation process of the proposed model. In the generation process, we specifically focus on the characteristics of the propagation DAG in social networks. Our framework allows for unified training across different scales and an arbitrary number of graphs, and it can generate graphs of a much larger scale (at least 10 times larger than the SOTA). Furthermore, without loss of generality, our model demonstrates strong extensibility to general graph generation tasks, which can be achieved by conveniently revising some modules. We will briefly explain this in the corresponding sections.

The key idea of our generation strategy lies in the directed iterative focus on the graph-level and edge-level generation processes, guided by G^{Φ}. The graph-level process primarily concentrates on the topology of the DAG. Then, utilizing the topological information, the edge level incorporates user attribute information via an attention mechanism to gauge the potential of forming an edge between a new user and existing users in the current graph context, thereby facilitating the generation of new directed edges. The newly formed edge structure is further integrated into the graph-level generation process. This allows for a sequential generation within the entire framework. Through this strategy, our approach maintains an orderly progression, ensuring consistent and coherent graph and edge generation. Referring to the scalable modeling process, our graph-level generating function and edge-level generating function are defined as follows.

$$h_i = f_G(h_{i-1}, \Omega(\varphi_{i-1})), \tag{3}$$

$$\varphi_i = f_E(h_i, F), \tag{4}$$

where h_i represents a vector encoding the state of the graph topology generated so far, φ_{i-1} is the predicted adjacency vector associated with the most recently generated user v_{i-1}, and $\varphi_{i-1}(j)$ $(j < i-1)$ signifies the probability of an edge existing between the most recently generated user v_{i-1} and the historical user v_j. $\Omega(x)$ is a one-hot decoder function to set the maximum value to 1 and set all other values to 0 from the vector x, indicating which historical user is most likely to form an edge with the newly introduced user in the condition of the current topological context. Next, we present the graph-level generative model f_G and edge-level generative model f_E, which pertain to the task of modeling the inherent DAG in the social network.

The generative model f_G we propose represents an integration of a Gated Recurrent Unit (GRU) (Chung et al. 2014) and an Exponential Variational Autoencoder (E-VAE), where E-VAE is predicated on sampling from an exponential distribution. Our approach employs a Gated Recurrent Unit (GRU) and a Variational Autoencoder (VAE) module. The GRU ensures temporal data generation continuity, while the VAE provides potent latent variable representation. Uniquely, we use an Exponential-VAE (E-VAE) for propagation Directed Acyclic Graphs (DAGs). The choice is the idea that the primary objective of the graph-level generation model is to effectively learn and assimilate the topological information inherent in propagation graphs. We therefore expect to use some of the conclusions obtained from the propagation data analysis as a priori guidance for graph generation. That is, we found a long-tailed distribution trend from the Cumulative Distribution Function (CDF) analysis of topological features. Further, we fitted these traits to various distributions using MLE, and found the exponential distribution consistently yielded the lowest MLEs. To further assess the extent to which the exponential distribution fits our data, we employ the skewness-based Kolmogorov-Smirnov (K-S) test. Some features, like structural virality, surpassed the 0.05 significance level in the K-S test against an exponential distribution, leading DAVA to adopt exponential sampling in the graph-level generation [3].

$$\log q_{\boldsymbol{\phi}}\Big(\mathbf{z} \mid \mathrm{GRU}(\mathbf{x}^{(k)})\Big) = \log \mathrm{Exp}\Big(\mathbf{z}; \boldsymbol{\lambda}^{(k)}\Big), \tag{5}$$

where $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathrm{GRU}(\mathbf{x}^{(k)}))$ represents the approximate posterior distribution of the latent variable $\mathbf{z}$ parameterized by $\boldsymbol{\phi}$, conditioned on the output of the GRU applied to the input data $\mathbf{x}^{(k)}$. Eq. (5) implies that the distribution of the latent variable $\mathbf{z}$ is inferred based on the information extracted from the input data by the GRU, which encapsulates the temporal dependencies in the input data and then is used to guide the generation of the latent space within the VAE framework. In this case, samples are drawn from an exponential distribution. Our goal is to model the latent variables $\mathbf{z}$ given the k-th observed data points $\mathbf{x}^{(k)}$ using an exponential distribution. Therefore, $\log \mathrm{Exp}(\mathbf{z}; \boldsymbol{\lambda}^{(k)})$ represents the natural logarithm of the probability density function of an exponen-

[3]The interpretability of E-VAE and the corresponding KL divergence are proven in the supplementary files.

tial distribution with a rate parameter $\boldsymbol{\lambda}^{(k)}$. Moreover, to ensure efficient backpropagation of gradients and avoid the potential issue of exploding exponentials within DAVA, we use a combination of specific transformations and the reparameterization trick.

Further, the generative model f_E is informed by the current network topology representation h_i, placing emphasis on the connection density between users with unique attributes under given topological contexts. The idea of dynamically adjusting the weights between different edges is worth noting for sparse propagation structures. Thanks to the distinct structural features of DAGs with user attributions, we are able to easily implement a lightweight, interpretable attention mechanism. For the $n\text{-}th$ historical user in the Φ sequence, we model it as a vector where the $n\text{-}th$ position is 1 and all other positions are 0. This vector signifies that the $n\text{-}th$ node will ultimately establish a connection with the new user, serving as the *value* in our attention mechanism. In a similar interpretive vein, we designate the new user's feature information as the *query*, while the historical users' feature information constitutes the *key*. To further refine our model, we aggregate the most recently updated global topological information, h_i, into the *query* and *key*. Lastly, based on the *query* and *key*, we are able to calculate the weight corresponding to each *value*. And multiplying the calculated weights by their corresponding *value* yields the probability of the existence of each edge, thereby quantifying the likelihood of the connection between the historical users and the current user. More specifically, the *query* and *key* are initially transformed through a linear transformation followed by an activation function:

$$Q_m = \sigma(W_q \cdot query_m + b_q), \tag{6}$$

$$K_n = \sigma(W_q \cdot key_n + b_q), \forall n \in (1, 2, ...i), \tag{7}$$

where $query_m = cat(F_m, h_i)$, $key_n = cat(F_n, h_i)$, W_q, b_q, W_k, and b_k represent the weights and biases in linear transformations of the *query* and *key*, respectively. Then, the scores are computed by taking the dot product between the transformed *query* and the transpose of the transformed *key*:

$$\text{scores}_{mn} = \text{Softmax}((Q_m^T \cdot K_n), \forall n \in (1, 2, ...i)). \tag{8}$$

Finally, the likelihood of the connection between the historical users n and the current user m is obtained by multiplying the weighted score with the corresponding *value*:

$$\text{likelihood}_{mn} = \text{scores}_{mn} \cdot value_n. \tag{9}$$

Loss Function Defination So far, the definition of the generative model is complete. To ensure effective training, a valuable loss function needs to be established. The definition of the objective function is as follows:

$$\text{loss} = \text{loss}_{\text{bce}} + \alpha * \text{loss}_{\text{KL}}, \tag{10}$$

where $\text{loss}_{\text{bce}} = -\left(\frac{1-\sum A_G}{|V|^2} y \log(\hat{y}) + \frac{\sum A_G}{|V|^2}(1-y)\log(1-\hat{y})\right)$ represents the reconstruction error loss during the generation process. Here a custom-weighted version is used to emphasize the importance of edges. A_G denotes the adjacency matrix corresponding to graph G^{Φ}, $\hat{y}$ is the predicted probability of the edge corresponding to φ, and y signifies the presence or absence of an edge in A_G. $\text{loss}_{\text{KL}} = \text{Norm}(\log\left(\frac{\mathbf{z}}{\lambda^{(k)}}\right) + \frac{\lambda^{(k)}}{\mathbf{z}} - 1)$ represents the normalized KL divergence between the latent representation and standard exponential distribution, $\mathbf{z}$ is the latent representation of the E-VAE corresponding to the graph-level generation.

Tricks for New Graph Generation In the new graph generation process, we ensure robustness from three aspects: the selection of new candidates, the achievable scale of generation, and the mapping of social phenomena in our models.

First, in order to ensure rigor in the sampling process during the generation of G', we construct a joint historical relationship network, or a union graph $\mathcal{G}$. Specifically, for any k propagation DAG $G_1, G_2, ..., G_k$ from the same platform and the same period, we merge nodes with identical unique user identifiers (UIDs) across these trees to fuse the k propagation DAGs into a single union graph $\mathcal{G}$. $\mathcal{G}$ serves as a foundational basis for sampling new users in the process of new graph generation.

Secondly, as the scale of the network increases, it becomes increasingly resource-intensive for the GRU to handle the growing input size corresponding to the expanded adjacency length of each node in G^{Φ} at every timestep. Fortunately, the proposed BFS permutation with node importance ensures stronger contextual information between a node v_i and its closer adjacent nodes in ordering Φ. Since under the ordering Φ, nodes slightly ahead or behind v_i can be either more or less crucial nodes at the same DAG depth of v_i, or nodes from the adjacent depth layers of v_i. This approach obviates the need to consider the far-distant relationship between v_i and the first node v_0 when $|V|$ is very large. Thus, the permutation strategy enables the effective utilization of a sliding window in two aspects. During the training phase, if the size of the training dataset $|V(G^{\Phi})|$ exceeds the sliding window size d (which concurrently serves as the input size for the GRU), we have the capacity to select the most recent d elements from each row within G^{Φ}. This effectively maps the data from a higher dimension $\mathbb{R}^{|V|\times|V|}$ to a lower-dimensional space $\mathbb{R}^{|V|\times|d|}$. During the new graph generation phase, for an expected graph with $|V'|$ nodes, where $|V'|$ exceeds the sliding window size d, generating node from d to $|V'|$ in f_G requires focusing solely on the information of the most recent d nodes. Moreover, our feature-driven attention mechanism is unaffected by the length of the *value*, eliminating the concern of $|V'|$ and $|V|$ in f_E.

The CEE phenomenon is inspired by analogous patterns observed generally in diverse fields such as advertising (Darke and Ritchie 2007) and communication (Metzger, Flanagin, and Medders 2010). The Credibility Erosion Effect (CEE) here refers to a gradual decline of credibility from the same person who repeatedly spreads and shares information. This effect has been widely observed in social networks (Turcotte et al. 2015), especially in the context of fake news and rumor diffusion, and regarding social media influencers. If a source consistently spreads questionable information, people may start doubting the credibil-

ity of this source or be influenced by other participants in this event. The effectiveness of incorporating this CEE phenomenon into the new graph generation process of DAVA has been successfully validated with real-world propagation data. To model this phenomenon, we introduce a decay mechanism, denoted as Ψ, to optimize the edge generation process $\Omega(\Psi(\varphi_i))$. After predicting the edge probabilities between a new user v_j and each historical user through f_E, these probabilities are adjusted. If a historical user v_i has k succeeding users, then the probability of an edge from v_i to v_j is reduced by a cumulative decay factor of β^k, where β is a decay factor very close to but less than 1. We have conducted extensive experiments to validate the effectiveness of these techniques.

Experiments

Datasets and Baselines

We used three datasets collected from two real-world social media platforms, Weibo and Twitter, for graph generation, namely Weibo (Ma, Gao, and Wong 2017), Twitter15, and Twitter16 (Liu et al. 2015; Ma et al. 2016). Based on the user IDs in 6,059 public propagation cascades from Twitter and Weibo, we collected the profiles of corresponding nearly 1 million distinct users, including verification status, number of tweets, registration date, number of fans, number of followings, and ratio of fans to followings. Through the extensive data collection, we formed joint historical relationship networks with user profiles of 480,405, 289,504, and 2,856,519, respectively. These networks serve as a priori knowledge for the graph generation task. The relevant information of the three datasets is shown in Tab. 1.

Statistic	Twitter15	Twitter16	Weibo
#users	480,987	289,675	2,856,741
#users in $\mathcal{G}$	480,405	289,504	2,856,519
#relations in $\mathcal{G}$	565,948	334,603	3,508,596
#tweets	1490	818	4664
#rumors	370	205	2244
#non-rumors	746	412	2082

Table 1: Statistics of the datasets. $\mathcal{G}$ is the largest component of the joint historical relationship network based on UIDs.

Comprehensive Evaluation of DAVA

Due to the absence of a unified standard for assessing the generation of propagation DAGs in the social network, a comprehensive extension of the evaluation metrics for DAVA has been undertaken from three facets. And we compare the SOTA methods of DAGG (Han et al. 2023), GVAE_MM (Zahirnia et al. 2022), D-VAE (Zhang et al. 2019), GraphVAE (Simonovsky and Komodakis 2018), GraphRNN (You et al. 2018b).

MMD Based Metric Evaluation First, the Maximum Mean Discrepancy (MMD) metric is widely used in the domain of graph generation, primarily assessing the similarity between two data distributions. A smaller value indicates a closer approximation. We follow and employ squared MMD between two sets of samples from distributions p and q based on the Reproducing Kernel Hilbert Space (RKHS) (Kawai, Mukuta, and Harada 2019), as shown in Eq. (11).

$$\begin{aligned}\mathrm{MMD}^2(p\|q) = \mathbb{E}_{x,y\sim p}[k(x,y)] + \mathbb{E}_{x,y\sim q}[k(x,y)] \\ - 2\mathbb{E}_{x\sim p,y\sim q}[k(x,y)],\end{aligned} \tag{11}$$

where k corresponds to the kernel function that operates on individual samples x and y drawn from the distributions.

Metrics	MMD^2			Time (h)		
Group	*T15*	*T16*	*Wb*	*100*	*1,000*	*10,000*
DAGG	0.287	0.244	0.291	0.417	-	-
GVAE_MM	0.316	0.263	0.242	0.25	0.667	-
D-VAE	0.271	0.216	0.255	1.167	-	-
GraphVAE	0.358	0.355	0.306	0.083	0.25	-
GraphRNN	0.331	0.357	0.401	0.133	0.333	-
DAVA	**0.149**	**0.134**	**0.201**	**0.002**	**0.01**	**0.5**

Table 2: The generation performance evaluation of different methods based on MMD metric. The time signifies the approximate hours needed for the model to generate a single graph of 100, 1,000, and 10,000 nodes, respectively. Symbol "-" indicates that the model could not successfully generate the graph of the corresponding scale using the maximum valid facility. The bold values represent the best results.

DAVA consistently outperforms all tested datasets, reducing the MMD by an average of 40% compared to the optimal D-VAE baseline and increasing the generation scale by two orders of magnitude. The superiority of more accurate generation can be attributed to four key factors: (1) Using statistical analysis from real-world data as prior knowledge, the E-VAE is more capable of representing the latent space of real-world topology structures. (2) The constructed joint relationship network provides historical prior knowledge of user relationships. (3) The attention mechanism dynamically focuses on edge connections based on the user feature representation. (4) We identify a phenomenon of CEE in social network propagation and effectively incorporate it into the generation process. The larger scale generation in less time is due to two reasons: (1) The BFS permutation with node importance creates a strong correlation for each node's context, allowing the use of a lightweight length sliding window for effective sequential generation without needing to concern the global context in the graph-level autoregressive process. (2) The interpretable attention mechanism allows for a lightweight module design, significantly reducing the various attention parameters compared to other attention models. In summary, DAVA generates propagation DAGs that are closer to reality, larger in scale, and quicker to produce.

Assessing the Realism of Generated Data Based on CCDF Second, we compare the characteristics of our generated data with those of real propagation data. Specifically, we examine topological characteristics such as breadth, depth, and structural virality, as well as the distribution of

attribute values of source and participating users across different intervals, in both rumor and non-rumor propagation DAGs $f_\theta(G)$. Due to the limited space, we present the most focused structural virality (Goel et al. 2016). Intuitively, assessing and comparing the difference in CCDF distributions in a visualization way between generative graphs with original real-world propagation data allows for a straightforward and convenient way to observe the goodness-of-fit of the generated data to the topological features or user attributes of real-world data from a statistical analysis perspective. As shown in Fig. 3, the CCDF curve of the DAVA-generated data closely matches that of the original data, demonstrating a high degree of similarity. By comparing these distributions, we visually assessed that DAVA has the capability to generate data most similar to the real propagation data, in an intuitive and straightforward manner.

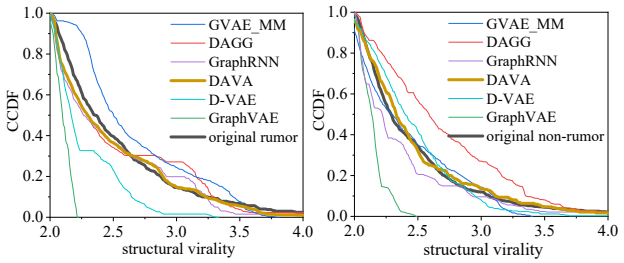

Figure 3: Comparation of the CCDF distribution differences in structural virality between graphs generated by different methods and real propagation data from Twitter15.

Utility of Generated Data in Downstream Tasks Third, numerous downstream tasks, such as influence maximization, fake news detection, information diffusion analysis, and source localization, rely on propagation models. We use source localization as an example to explore if using generated data can enhance model predictive ability in real scenarios. Here, two localization models, GCNSI (Dong et al. 2019) and TGASI (Hou et al. 2023), are used. In the experiment, the original groups train localization models on 9/10 of the Twitter propagation data and test on the remaining 1/10. For comparison, the augmentation groups additionally generate 1,000 real-world propagation graphs by DAVA or SOTAs for training, while the control groups simulate 1,000 snapshots based on SI, SIR, IC, and LT models. Tab. 3 shows that the use of simulation data from traditional propagation models leads to a decrease in the performance of downstream tasks in real-world propagation scenarios, suggesting that these models may have limited relevance for real-world tasks. Conversely, augmenting with real generated data improves outcomes, particularly improving most when using propagation data generated by DAVA, thereby emphasizing the importance of graph generation and DAVA's utility.

Ablation Study

We further investigate the impact of each module in DAVA on the performance of the graph generation to demonstrate their necessity. The critical modules of DAVA include the E-VAE, user relationship attention, loss function, and CEE-based decay mechanism. So some variants of DAVA are developed to compare with DAVA. DAVA_Norm uses the normal distribution to represent the latent space of the graph-level generation in Eq. (5). DAVA_Att uses the attention module in Transformer (Vaswani et al. 2017) to replace the proposed interpretable attention in Eqs. (6)-(9). DAVA_GRU uses the autoregressive generative model to replace the proposed attention in Eqs. (6)-(9). DAVA_EN replaces the unique loss function in Eq. (10) with the binary cross-entropy loss function. DAVA_CEE− removes the decay mechanism. Due to the limited space, we only present the variants of DAVA in the Twitter16 dataset. As shown in Tab. 4, it would lead to a generation similarity decrease, a generation scale reduction, or a generation time increase, no matter removing or replacing critical modules.

Strategy	Original	Augmented (DAVA/SOTA)	Control
GCNSI	0.532	**0.613**/0.582	0.512
TGASI	0.787	**0.825**/0.808	0.755

Table 3: Source detection accuracy of localization methods under different groups of training sets.

	MMD^2	Graph Scale	Time (h)
DAVA	**0.134**	**x10^4**	**0.001/0.01/0.5**
DAVA_Norm	0.207	x10^4	0.001/0.01/0.5
DAVA_Att	0.151	x10^3	0.005/0.083/-
DAVA_GRU	0.322	x10^3	0.017/0.217/-
DAVA_EN	0.236	x10^4	0.001/0.01/0.5
DAVA_CEE−	0.181	x10^4	0.001/0.01/0.5

Table 4: The generation performance evaluation of the variant model from DAVA based on MMD metric in Twitter16.

Conclusion

In this paper, we generate large-scale and diverse social media propagation graphs by incorporating user attributes from Twitter and Weibo. Our analysis of nearly a million users' social attributes, focusing on propagation characteristics and user features, revealed a prevalent exponential distribution and the presence of a credibility erosion effect in these media. Leveraging these prior knowledge, we develop a DAVA model to enhance the realism of generated data in a low-cost way. In the future, we are committed to collecting more propagation data and generating larger-scale graphs.

Acknowledgements

This work was supported by the National Key R&D Program (no. 2022YFE0112300); the National Natural Science Foundation for Distinguished Young Scholars (no. 62025602); the National Natural Science Foundation of China (nos. U22B2036, 62261136549, 11931015 and 61976181); the Fok Ying-Tong Education Foundation, China (Grant No. 171105); and the Tencent Foundation and XPLORER PRIZE.

References

Allport, G. W.; and Postman, L. 1947. The psychology of rumor. *Russell & Russell*.

Chen, W.; Castillo, C.; and Lakshmanan, L. V. 2022. *Information and influence propagation in social networks*. Springer Nature.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, 1–9.

Darke, P. R.; and Ritchie, R. J. 2007. The defensive consumer: Advertising deception, defensive processing, and distrust. *Journal of Marketing research*, 44(1): 114–127.

Dong, M.; Zheng, B.; Quoc Viet Hung, N.; Su, H.; and Li, G. 2019. Multiple Rumor Source Detection with Graph Convolutional Networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 569–578.

Goel, S.; Anderson, A.; Hofman, J.; and Watts, D. J. 2016. The structural virality of online diffusion. *Management Science*, 62(1): 180–196.

Guille, A.; Hacid, H.; Favre, C.; and Zighed, D. A. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2): 17–28.

Han, X.; Chen, X.; Ruiz, F. J.; and Liu, L.-P. 2023. Fitting Autoregressive Graph Generative Models through Maximum Likelihood Estimation. *Journal of Machine Learning Research*, 24(97): 1–30.

Hou, D.; Wang, Z.; Gao, C.; and Li, X. 2023. Sequential attention source identification based on feature representation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4794–4802.

Hou, D.; Yin, S.; Gao, C.; Li, X.; and Wang, Z. 2024. Propagation Dynamics of Rumor vs. Non-rumor across Multiple Social Media Platforms Driven by User Characteristics. arXiv:2401.17840.

Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-BERT: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 459–469.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332. PMLR.

Kawai, W.; Mukuta, Y.; and Harada, T. 2019. Scalable generative models for graphs with graph attention mechanism. *arXiv preprint arXiv:1906.01861*.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429.

Li, Q.; Hu, B.; Xu, W.; and Xiao, Y. 2022. A group behavior prediction model based on sparse representation and complex message interactions. *Information Sciences*, 601: 224–241.

Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; and Battaglia, P. 2018. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*.

Ling, C.; Jiang, J.; Wang, J.; and Liang, Z. 2022. Source localization of graph diffusion via variational autoencoders for graph inverse problems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1010–1020.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015. Real-time Rumor Debunking on Twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1867–1870.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Meeyoung, C. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *The 25th International Joint Conference on Artificial Intelligence*. AAAI.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 708–717.

Metzger, M. J.; Flanagin, A. J.; and Medders, R. B. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 60(3): 413–439.

Shakarian, P.; Bhatnagar, A.; Aleali, A.; Shaabani, E.; Guo, R.; Shakarian, P.; Bhatnagar, A.; Aleali, A.; Shaabani, E.; and Guo, R. 2015. The independent cascade and linear threshold models. *Diffusion in Social Networks*, 35–48.

Simonovsky, M.; and Komodakis, N. 2018. Graphvae: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, 412–422. Springer.

Turcotte, J.; York, C.; Irving, J.; Scholl, R. M.; and Pingree, R. J. 2015. News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of computer-mediated communication*, 20(5): 520–535.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. *International Conference on Learning Representations*, 1–12.

Vinyals, O.; Bengio, S.; and Kudlur, M. 2015. Order matters: Sequence to sequence for sets. *International Conference on Learning Representations*, 1–11.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Wang, J.; Jiang, J.; and Zhao, L. 2022. An Invertible Graph Diffusion Neural Network for Source Localization. In *Proceedings of the ACM Web Conference*, 1058–1069.

Wang, Z.; Hou, D.; Gao, C.; Huang, J.; and Xuan, Q. 2022. A rapid source localization method in the early stage of large-scale network propagation. In *Proceedings of the ACM web conference 2022*, 1372–1380.

Wang, Z.; Hou, D.; Gao, C.; Li, X.; and Li, X. 2023. Lightweight source localization for large-scale social networks. In *Proceedings of the ACM Web Conference 2023*, 286–294.

Welling, M.; and Kipf, T. N. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 1–14.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xia, W.; Li, Y.; Wu, J.; and Li, S. 2021. DeepIS: Susceptibility estimation on social networks. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 761–769.

You, J.; Liu, B.; Ying, Z.; Pande, V.; and Leskovec, J. 2018a. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31.

You, J.; Ying, R.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018b. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, 5708–5717. PMLR.

Zahirnia, K.; Schulte, O.; Naddaf, P.; and Li, K. 2022. Micro and macro level graph modeling for graph variational auto-encoders. In *Advances in Neural Information Processing Systems*, volume 35, 30347–30361.

Zhang, M.; Jiang, S.; Cui, Z.; Garnett, R.; and Chen, Y. 2019. D-VAE: A Variational Autoencoder for Directed Acyclic Graphs. In *Advances in Neural Information Processing Systems*, 1586–1598.