PRIM: Encoding Propagation Probability and Role-Aware Representation for Influence Maximization

Niran Deng[1], Jiuyang Tang[1(✉)], Yang Fang[2], Tianyang Shao[2], Jinzhi Liao[1], and Xiang Zhao[1]

[1] Laboratory for Big Data and Decision, National University of Defense Technology, Changsha, China
{dengniran,jiuyang_tang,liaojinzhil2,xiangzhao}@nudt.edu.cn
[2] National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, China
{fangyang12,shaotianyang}@nudt.edu.cn

Abstract. Influence maximization in networks has demonstrated robust efficacy in the field of identifying a set of initial influential key nodes. Although recent machine learning based methods have enhanced the performance in unknown graphs for its stronger generalization, they ignored the diversified underlying information diffusion patterns in real world. Moreover, there exist the strong coupling characteristics between the network topology structure and the individual nodes when information spreads. To address the aforementioned issues, we propose a novel framework, i.e., Encoding Propagation Probability and Role-Aware Representation for Influence Maximization (PRIM). Specifically, we utilize a feature extractor module to distinguish the nodes with superior information transmission ability, which includes a graph diffusion convolution to learn multi-hop propagation probability matrix and a graph convolutional network to learn the role structural embedding. Then we input the features into a multi-head attention message passing module to rank the influential nodes. Extensive experiments are conducted which validate the effectiveness of PRIM and its superiority to state-of-the-art methods.

Keywords: Influence maximization · Information diffusion · Propagation probability · Structural role extraction · Social networks

1 Introduction

The social network is a structure comprised of individuals and social interactions between users. The **influence maximization** (IM) problem stands out as the most critical algorithmic challenge in the study of influence spread in social networks, which aims at figuring out a set of influential nodes with significant spreading capabilities. And this issue has been the subject of extensive researches over the past two decades due to its significant importance in various domains including its substantial value in multiple aspects such as political campaign [2], social media marketing [11,29], rumor control [7], and social recommendations [32].

During the past twenty years, researchers have put forward heuristic algorithms, approximation algorithms, and algorithms based on machine learning (ML) step by step [21,27]. Heuristic algorithms and approximation algorithms are time-consuming and do not have good generalization ability. When the social networks dynamically change in practical applications, they need to re-search and adjust the heuristic rules through trial and error to get better results. ML-based algorithms include deep reinforcement learning (DRL) [8,19] and graph neural networks(GNNs) [17,18,22]. DRL-based methods find the best action sequences as the optimal seed nodes by calculating and ranking the cumulative rewards. GNNs-based algorithms learn the node representations to measure and rank the importance of each node to network topological stability and select the top-ranked ones as the seed nodes. They naturally surpass the traditional algorithms because of their advantages of solving the discrete problem in continuous space in an end-to-end manner.

Whereas, these models still meet some limitations: (a) They ignore the intuitive structural features of important influential nodes in network topology, e.g., nodes with higher degrees typically relaying information faster and treating each node identically can impede training efficiency. (b) They pay less attention to how the network structure affects the information dissemination, which is key to observe the actual underlying propagation process of the network.

Inspired by the successes of GNNs-based approaches applied to IM problem [17,18,22], we want to solve the above two issues from two aspects: (a) Influence propagation characteristics: by using GNNs module we can module the propagation process in a more efficient way. (b) Influencer characteristics: we should focus on the most influential and non-influential nodes and the role-aware features extractor module can well exploit the graph topology structure to distinguish the nodes' message passing competence in the spreading process.

In this paper, we propose an unsupervised learning framework for IM problem, encoding Propagation Probability and Role-Aware Representation for Influence Maximization, called PRIM, which mainly focuses on nodes with superior information transmission capacity by learning the representations of both the global graph structure and the individual spreading ability of each single node. Specifically, first, we encode the global influence propagation probability of each node using Graph Diffusion Convolution (GDC) and the nodes' role-aware representations using Graph Convolutional Network (GCN). Second, to focus on nodes with greater information spreading capacity, a multi-head attention message passing module is employed to automatically allocate varying attention weights to different nodes. Finally, after training the model on a small quantity of samples, we infer the optimal seed sets on different sizes of the sampled datasets. Experiment results on real networks validate the superiority of our PRIM over the state-of-the-art competitions in terms of overall diffusion count under the weighted the independent cascade (IC) model in most cases.

The major contributions of this paper are summarized as follows:

1. We propose a novel framework named PRIM to solve the IM problem, which uses the GDC module to learn the propagation characteristics from multi-hop neighbors and the GCN module to learn the global topology structure representation.
2. To the best of our knowledge, this work represents the first attempt to model the underlying propagation process in a GNN-based manner, through combining propagation features and network topology features. We capture the nodes with higher information dissemination competence by a multi-head attention message passing module.
3. We conducted extensive experiments on some datasets, which justified our model's advantage not only over the state-of-the-art IM methods but also over some other node importance evaluation metrics.

The rest of the paper is organized as follows: Section 2 introduces some related works on IM methods and information diffusion. Next, in Section 3 we formalize the problem and present our PRIM framework. Experiment results and analysis are displayed in Section 4. Finally, Section 5 shows our conclusion of our work.

2 Related Work

Related prior work includes studies of influence maximization and information diffusion.

2.1 Influence Maximization

Influence Maximization (IM) is a typical combinatorial optimization problem, which is firstly introduced by Kempe et al. [16]. IM problem aims to select a size-k seed set so that the expected number of influence spread can be maximized given a diffusion model such as the linear threshold (LT) model, IC model, and the susceptible-infectious-susceptible (SIS) epidemic model. OPIM [28] and SubSIM [14] proposed algorithms mainly focusing on improving the empirical efficiency. The main challenge of these traditional algorithms not only lies in the NP-hardness of the IM problem and #P-hardness of the accurate influence diffusion estimation [9], but also in getting the practical diffusion models and its corresponding parameters in real-world social networks, since they are given some specific models. Furthermore, it is difficult to meet the demand for algorithm performances due to the growing network scales. Thus it is almost impossible to further optimize and improve their performances.

As with the rapid progress of artificial intelligence, methods based on deep learning are widely used in solving IM problem due to the advantages of fast solving speed and high generalization ability [20]. IMINFECTOR [25] uses the logs of diffusion cascades to embed the diffusion probabilities and then used these probabilities to find a seed set using a greedy approach. [17] uses the struc2vec to embed every node in the network and puts them into a GNN-based regressor to rank nodes on the basis of predicted influence under the SIR and IC model. Similarly, GLSTM [18] feeds the generated node centrality feature vectors into a graph-based LSTM model to predict the probable influence of every node in the target network. ToupleGDD [8] incorporates three coupled GNNs for network embedding and double deep Q-network (DDQN) technique for parameter learning. DeepIM [22] employs a framework to learn the diversified information diffusion patterns in an end-to-end and data-driven manner.

2.2 Information Diffusion

Modeling the information diffusion plays a significant role in finding the optimal seed set in the IM problem. However the pre-defined diffusion models such as LT, IC [16] and SIS [22] models rely on simplified assumptions and are often constrained to specific scenarios, thereby limiting their real-world applicability. Besides, they ignore the users' information diffusion patterns and the interactions between network topology and users. To fully capture the complex interaction parameters of underlying propagation models, deep learning-based methods have garnered significant attention in recent years. DEEPIS [31] uses the power sequence of the influence probability matrix and a two-layer GNNs to regress the susceptibility of each node. Subsequently, the estimation propagates in the neighbors based on the IC probability. IVGD [30] designs a flexible architecture of reversible graph diffusion models to autonomously learn the inherent rules of the disseminating process. DCRS [33] proposes a graph diffusion neural network which mimics the process of information diffusion for competence encoding. GIN-SD [10] combines the propagation information and user states into the feature vectors of users to capture the higher influential nodes. These methods illustrate that another important concern of the IM problem is highly related to the structural characteristics of the network. And thus different end-to-end manners are introduced to solve this problem.

3 Methodology

3.1 Problem Definition

IM Problem For the IM problem, the social network can be denoted as $G = (V, E)$, in which the user set is V and E describes the social relationship between users. Given the fixed seed nodes size, the objective of the IM problem is to strategically choose the optimal seed nodes to maximize the diffusion scale.

$$\mathbf{S}^* = \arg\max M(\mathbf{S}; G, R, P), \tag{1}$$

where R refers the role structure of G and P refers to the diffusion probability of G. $\mathbf{S}^*$ represents the expected optimal seed nodes which enables the diffusion model $M(\cdot)$ maximal.

WIC model In the social network G, for each edge $\langle u, v\rangle \in E$, u and v are neighbors of each other. Suppose node u gets activated at timestamp i , then

u has a single chance to activate its inactive neighbor v with the propagation probability $p(u, v) \in [0, 1]$ at timestamp $i+1$. Once node v is triggered, it remains active in the next timestamps. Finally, after the last timestamp, node u cannot activate any of its neighbors and the diffusion process comes to an end.

$$p_{u,v} = 1/d_v^{in} \tag{2}$$

where d_v^{in} denotes the in-degree of node v.

3.2 Model Framework

The model framework is depicted in Fig. 1. It contains three parts: feature extractor module, multi-head attention message passing module and optimal seeds inference module. In feature extractor module, we concatenate user state embedding, influence probability embedding and role structure embedding together to learn the users' representations in the process of information dissemination. To pay more attention to the users with greater transmission capacities, the multi-head attention message passing module is used to distinguish the influential nodes and non-influential nodes. After the model training stage, we resample and rank all the sampled seed nodes to acquire the most influential nodes as the optimal seed nodes.

3.3 Feature Extractor Module

In social networks, information spreading can be affected by lots of interacting factors, including network topology structure, users' activities, information characteristics and so on. These entwined factors contribute to the diversity of diffusion patterns in the real world. To better model the information diffusion process, we extract the user state features, influence probability features and role structure features to illustrate the essence of information spreading.

User State Embedding Initially a fixed rate of nodes S are sampled to be the seed nodes which received the information and began to spread it to its neighbors. Different initialized seed nodes may lead to different diffusion paths. Therefore, we encode the user's initial state to better learn the information of seed nodes after propagation . For network G, the user state feature $\mathbf{X}_i^1$ can be determined by the following rules:

$$\mathbf{X}_i^1 = \begin{cases} 1, & v_i \in S \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

in which the user states can be categorized into two sets.

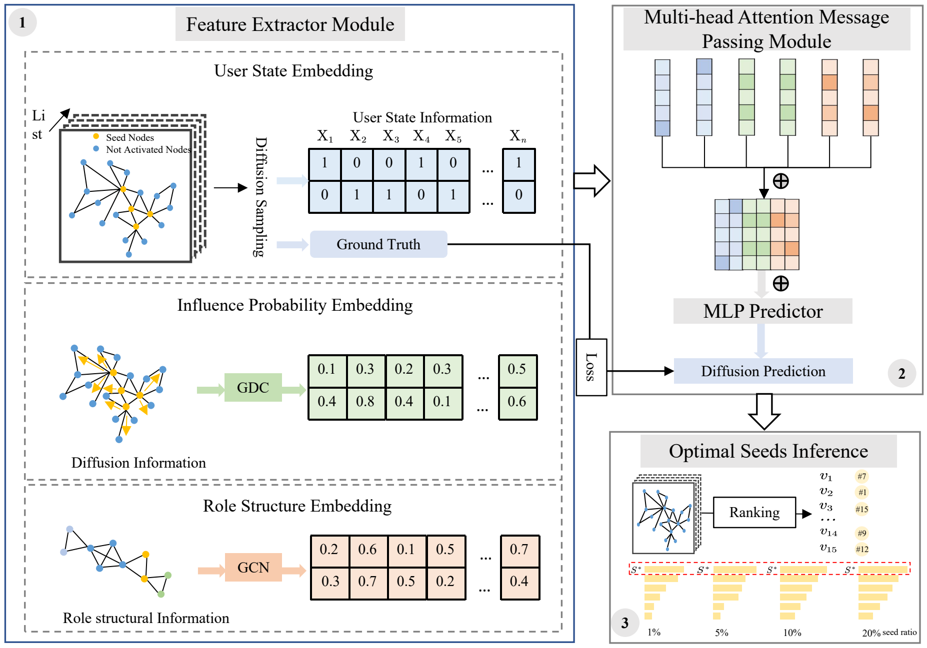


Fig. 1. The framework of our proposed PRIM consists of feature extractor module, multi-head attention message passing module and optimal seeds inference module. The feature extractor module learns the network topological structures from the perspective of propagation characteristics and influential nodes' features. With these three embeddings, the information diffusion representation is learned by way of GNNs. Multi-head attention message passing module is trained aiming at identifying the influential nodes. And then resample the test sets of different sizes to select the top-k seed nodes by ranking them using the well-trained multi-head attention message passing module.

Influence Probability Embedding According to [34], the diffusion probability of users along a social connections abides by a power-law relationship with the in-degree and out-degree of the nodes. Different from random walk, a node can only be influenced once in the information spread stage. Though the diffusion models like LT, IC, SIS could generate some realistic diffusion trees and simulate certain features of the spreading process, they can not connect the propensity of social ties with the spreading characteristics of the involved users due to the unavailability of underlying latent social topology. Therefore, we use the graph diffusion convolution module to model the diffusion probability of users.

For the underlying social graph structure G, the influence probability matrix P in one-hop can be defined as Eq.4.

$$P_{n\times n} = (p_{u,v})_{n\times n}, , \tag{4}$$

where $p_{u,v}$ represents the activation possibility of node u to node v in timestamp $t+1$ once node u is activated in timestamp t. $p_{u,v}$ is calculated as the reciprocal number of degrees of node v.

However, this only takes the one-hop neighbors of the node i into account. In fact, information spreading usually reaches high-order neighbors. GDC [13] is the spatial localization extension of message passing in GNNs. It allows GNNs to learn from multi-hop messages and is no longer limited to the first-hop neighborhood. Therefore, note that the P' is global influence probability matrix involving the multi-hop message passing among users.

$$P' = \sum_{k=0}^{\infty} \theta_k T^k, \tag{5}$$

where θ_k is the coefficient, T is the transfer matrix defined by PD^{-1}, P is the influence propagation probability matrix, and D is the diagonal degree matrix, $d_{u,u} = \sum_j p_{u,v}$.

Then we calculate the graph's diffusion probability at d timestamp, which is defined as d-step influence probability [31]. It observes the cascading paths at the length of d from the source seed nodes to their neighbor nodes. Thus the influence probability embedding can be represented as Eq.6:

$$\mathbf{X}_i^2 = [x_i, P^T x_i, (P^T)^2 x_i ..., (P^T)^d x_i], \tag{6}$$

Role Structure Embedding In social networks, certain nodes play the role of building up global communication capacity while some just perform as the periphery-nodes. By dividing all the nodes into several kinds of roles and encoding each node's role representation, we could capture the structural similarities features of nodes and mine the correlations among node features. These features are also key factors of the propagation patterns.

Firstly we construct the role graph by applying the RoIX algorithm [15] based on the role embedding $\mathbf{R}$, which recursively aggregates the ego-nets' features of the original graphs, such as its degree, cluster coefficient, percentage of edges within and leaving the ego-nets, into a feature matrix $\mathbf{X}$. And then we use non-negative matrix factorization (NMF) to automatically figure out a rank r approximation $\hat{\mathbf{R}} \times \mathbf{W} \approx \mathbf{X}$:

$$\min_{\hat{\mathbf{R}}, \mathbf{W}} \|\mathbf{X} - \hat{\mathbf{R}}\mathbf{W}\|_X^2, \quad \text{s.t. } \hat{\mathbf{R}}, \mathrm{W} \geq 0, \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{r \times d}$ is the role contribution matrix illustrating every role's correlation to local structural features. $\hat{\mathbf{R}} \in \mathbb{R}^{N \times r}$ is represented as the role embedding matrix which describes every node's mixed-membership probability belonging to the r predefined roles. Here r depends on the minimum description length criterion. $\| \cdot \|_X$ is the is the Frobenius norm.

Secondly, based on the $\hat{\mathbf{R}} \in \mathbb{R}^{N \times r}$, the role graph $\mathcal{G}_r$ could be constructed, which reconnects the nodes with similar regional structure features and captures the importance of roles on the network structure. And the role similarity between the role embeddings of node pairs is worthy of noting. The role similarity r_{ij} of node u_i and u_j is calculated by Eq.8:

$$r_{i,j} = \frac{\langle \hat{\mathbf{R}}_i, \hat{\mathbf{R}}_j \rangle}{\|\hat{\mathbf{R}}_i\| \|\hat{\mathbf{R}}_j\|}, \tag{8}$$

where $\hat{\mathbf{R}}$ represents the role embedding of every node, $\langle \cdot, \cdot \rangle$ represents the dot product of two node vectors. According to $r_{i,j}$ between nodes, edge links would be assigned to top k node pairs with similar roles, which generates the role graph $\mathcal{G}_r$.

Lastly, for role graph $\mathcal{G}_r$, $\mathbf{A}_r$ is the adjacency matrix, and $\mathbf{D}_r$ is the diagonal degree matrix. We use a GCN to encode the role structure embedding $\mathbf{X}_i^3$ of the network, in which one GCN layer is described in Eq.9.

$$\mathrm{Z}^{(l-1)} = \mathrm{ReLU}(\tilde{\mathrm{D}}_r^{-\frac{1}{2}} \tilde{\mathrm{A}}_r \tilde{\mathrm{D}}_r^{-\frac{1}{2}} \mathrm{Z}^{(l-2)} \mathrm{W}_r^{(l-2)}), \tag{9}$$

At the last layer, we calculate the role structure scores of each node as follows, which is used to be the third feature of nodes V:

$$\mathbf{X}_i^3 = \mathrm{D}_r^{-\frac{1}{2}} \tilde{\mathrm{A}}_r \tilde{\mathrm{D}}_r^{-\frac{1}{2}} \mathrm{Z}^{(l-1)} \mathrm{W}_r, \tag{10}$$

where $\mathrm{Z}^{(l-1)}$ denotes the node features at $(l-1)$-layer, $\sigma(\cdot)$ stands for the activation function, W_r denotes the learnable parameter matrix at each layer. By stacking L-layer GCNs, we could obtain the final role structure matrix $\mathbf{X}^3$.

Feature Concatenation Lastly, we adopt the operation of concatenation to amalgamate the diverse user feature representations, culminating in the derivation of the ultimate user embedding vector as follows:

$$\mathbf{X}_i = \mathbf{X}_i^1 \oplus \mathbf{X}_i^2 \oplus \mathbf{X}_i^3, \tag{11}$$

where $\oplus$ stands for the concatenation operation.

Then, we concatenate all node representations $\mathbf{X}_i$ by vertically stacking them into the matrix $\mathbf{X}$.

$$\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \cdots; \mathbf{X}_N]. \tag{12}$$

Features mining as above promotes the exploration of richer network topology structure information representations, thereby contributing to the propagation modeling and higher efficacy and stability of finding the optimal seed nodes under the WIC model.

3.4 Multi-head Attention Message Passing Module

The cascading effect represents that the activation of a node in a network triggers the subsequent activation of its neighboring nodes, thereby forming a diffusion cascade on social networks. This is consistent with the message-passing mechanism in GNNs [8], which is widely used to model geometric network structures.

By iteratively updating node representations through aggregating information from neighbor nodes, they bear resemblances to the influence diffusion model. Taking the difference of information transmission efficiency among nodes into consideration, we concentrate on the nodes with higher information transmission capability. By employing the multi-head attention mechanism, the diverse feature representations generated above can be captured. Let $\mathbf{H} \in \mathbb{R}^{n\times d}$ be the node feature matrix and $\mathbf{Z} \in \mathbb{R}^{n\times d}$ the output status matrix of final information diffusion. We can define a global attention message passing as follows:

$$\mathbf{Z} = \mathrm{softmax}(\frac{\mathbf{HW}_Q(\mathbf{HW}_K)^\top}{\sqrt{d}})\mathbf{XW}_V, \tag{13}$$

where $\mathbf{W}_Q$,$\mathbf{W}_K$ and $\mathbf{W}_V$ are learnable projection matrices.

Then we evaluate the diffusion competence of each node v_i by using MLP to regress the matrix $\mathbf{Z}$ to the prediction of information diffusion influence as Eq.14, and the loss function can be calculated by Mean Absolute Error (MAE), defined as Eq.16.

$$\mathbf{Z}_i^{score} = \mathrm{ReLU}\left(\mathbf{W}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{b}\right), \tag{14}$$

$$\hat{y} = \sum_{i=1}^{n} \mathbf{Z}_i^{score}, \tag{15}$$

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - \hat{y}_i|. \tag{16}$$

where W is the learnable model parameter. $\mathbf{Z}_i^{score}$ can be interpreted as the probability that the node v_i belongs to the optimal seed set. Finally, We select the top-k nodes with the highest influence scores to form the target seed set $\mathbf{S}^*$.

4 Experiments

4.1 Datasets and Baselines

Our proposed PRIM model is compared with some other methods upon the following datasets often used in IM problem, as are depicted in Table 1.

Jazz [26]: This dataset constitutes a collaboration network of jazz musicians, wherein each node signifies an individual musician and each edge denotes a pair of musicians who have performed together in a band.

Cora ML [23]: This network encompasses a corpus of computer science research papers, where each node represents an individual paper and each edge signifies a citation relationship between two papers, indicating that one paper cites the other.

PPI [6]: This network represents the type and intensity of interaction between coding genes. Each node stands for the protein and each edge describes

Table 1. Characteristics of the datasets.

Network	|V|	|E|	#Avg.D
Jazz	198	2742	27.69
Cora ML	2,810	7,981	5.68
PPI	2,224	6,609	5.94

the interaction between proteins, which are key determinants of protein function in biological systems.

We compared PRIM with the following node centrality-based methods and IM approaches including traditional IM and ML-based IM to verify our PRIM's advantages.

Degree Centrality (DC) [1]: This ranks the nodes in the order of node degrees and select the top-k nodes as the optimal seed nodes.

Closeness Centrality (CC) [3]: It describes how the node is close to other nodes and select the top-k ones as the optimal seed nodes. It can be calculated as the reciprocal of the average shortest path distance between the targeted node and other nodes.

Betweenness Centrality (BC) [12]: This metric calculates the frequency of all-pairs shortest-paths through a node and selects the top-k ones as the optimal seed nodes.

Harmonic Centrality (HC) [4]: This metric is a variant of Closeness Centrality and emphasizes the ability of a node to act as a connection point in the network.

Eigenvector Centrality (EC) [5]: This is based on the eigenvectors of the adjacency matrix. It asserts that the node's importance is measured by its neighbors' importance.

Collective Influence (CI) [24]: It incorporates the information about nodes' influence at the global level and select the top-k ones as the optimal seed nodes.

OPIM [28]: The node has the flexibility to pause the algorithm at any point in time to request a solution to the IM problem, along with its associated approximation guarantee.

SubSIM [14]: It is an efficient random RR set generation algorithm under IC model.

ToupleGDD [8]: It incorporates three coupled GNNs for network embedding and double DQN for parameters learning to solve the IM problem.

DeepIM [22]: It proposes to learn the various information diffusion patterns by way of generatively characterizing the latent representation of seed sets under the Variational Autoencoder (VAE) and GNNs models.

4.2 Experiment Settings

All the experiments were conducted in PyTorch on a workstation with GPU of GeForce GTX 3090 Ti. For the baseline models, we set hyper-parameters as are stated in their original instructions and fine-tune them on each dataset. For our

PRIM model, the attention heads' number was set to 12 with hidden size 64 for all datasets except Jazz which is set to 32. The multi-layer perception (MLP) layer was set to 4. We chose Adam with learning rates $1e-3$, $1e-4$ and $1e-5$ for optimizing.

4.3 Results and Analysis

In this subsection, we present the results of our comparative experiments and offer the insights based on these findings.

Comparison with State-of-the-art Methods To validate the effectiveness of PRIM, we conduct comparisons with other approaches on some datasets in maximizing the influence under a weighted cascade version of the IC model. We selected 1%, 5%, 10%, and 20% nodes in each dataset as seed nodes. The weighted IC model was allowed to run until the diffusion process ceases, and we recorded the average influence spread over 100 simulation rounds as the IM performances. The results are reported as the percentage of ultimately infected nodes, calculated as the ratio of the number of infected nodes to the total number of nodes in the network. As is shown in Table 2, in the analysis of the experimental results, we identify several key observations.

Table 2. Results (diffusion count %) of different models on Cora ML, Power Grid and Jazz datasets under weighted IC model. The best is marked in **bold blue** while the second best is in <u>underline</u>.

Dataset / Model	Cora ML				PPI				Jazz			
	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
DC	<u>15.6</u>	**28.5**	38.3	49.4	42.3	51.7	60.8	72.1	**8.7**	26.5	38.8	48.5
CC	14.1	25.3	31.9	43.1	41.8	49.8	57.9	69.4	**8.7**	26.0	35.5	47.5
BC	15.0	<u>28.2</u>	39.1	51.7	42.5	54.8	63.5	<u>75.9</u>	**8.7**	28.1	36.3	<u>50.6</u>
HC	15.0	25.4	32.2	43.2	43.1	51.7	58.5	69.3	**8.7**	<u>28.8</u>	36.5	49.9
EC	8.9	16.1	24.6	37.5	41.7	46.8	55.3	65.8	<u>5.7</u>	23.8	34.6	45.9
CI	15.2	27.5	34.8	45.2	<u>43.9</u>	52.8	60.5	71.5	**8.7**	25.8	40.5	49.0
OPIM	13.4	26.9	37.4	50.9	43.1	56.3	62.3	72.7	2.4	20.1	34.4	46.8
SubSIM	10.1	25.7	36.8	51.1	43.5	54.1	**65.0**	74.5	3.6	18.8	37.6	44.7
ToupleGDD	10.6	27.5	38.5	51.5	42.3	<u>55.3</u>	<u>64.8</u>	**77.2**	3.3	20.4	37.2	45.7
DeepIM	14.1	28.1	<u>39.6</u>	<u>52.4</u>	42.4	49.8	59.8	72.1	4.9	23.3	<u>41.5</u>	49.9
PRIM	**15.7**	**28.5**	**39.9**	**53.1**	**47.8**	**57.1**	**65.0**	74.6	**8.7**	**29.7**	**41.9**	**55.2**

First, across these datasets, our PRIM consistently outperforms other methods in most cases. Second, node centrality-based methods occasionally achieve top performance, particularly on the Jazz dataset, where they rank first or second. This superior performance likely stems from two factors: (1) Jazz has the highest average degree among all networks, and (2) it is the smallest dataset in terms of network size. However, their accuracy diminishes on larger datasets, primarily due to the lack of propagation feature modeling. Third, the traditional benchmark methods OPIM and SubSIM, which rely on reverse-set sampling, show significantly lower performance than PRIM on Cora ML and Jazz datasets. An exception occurs on the PPI dataset with 10% seed nodes, where SubSIM matches PRIM's performance. These two methods rely on the approximation techniques and thus are not so accurate as PRIM. Finally, machine learning-based approaches (ToupleGDD and DeepIM) demonstrate competitive performance against traditional models, benefiting from enhanced scalability and generalization capabilities. However, these methods fail to explicitly model information diffusion dynamics. Furthermore, their network topology analysis neglects the heterogeneous roles nodes play during propagation. This leads to inaccurate modeling of the underlying end-to-end diffusion pattern. In contrast, PRIM introduces a robust framework that jointly learns end-to-end diffusion patterns and identifies high-influence node sets. This approach not only better captures latent diffusion dynamics but also addresses scalability constraints.

Table 2 reveals variability in second-best performers across networks, reinforcing that all methods provide approximate solutions to the NP-hard influence maximization problem.

Ablation Study and Analysis To validate PRIM's design, we conduct ablation studies on its key components (Table 3). The results demonstrate that: (1) both influence probability embedding and role structure embedding contribute significantly to performance, and (2) PRIM's integrated implementation outperforms individual modules, validating the synergy between these components.

1. PRIM w/o R removes the module of role structure embedding and the concatenated embedding is changed into $\mathbf{X}_i = \mathbf{X}_i^1 \oplus \mathbf{X}_i^2$.
2. PRIM w/o P removes the influence probability embedding and the concatenated embedding is changed into $\mathbf{X}_i = \mathbf{X}_i^1 \oplus \mathbf{X}_i^3$.

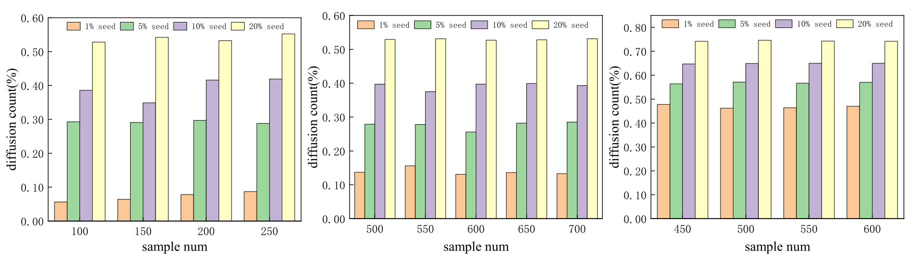
Influence Probability Embedding We validate the significance of the influence probability embedding through evaluating the variant PRIM w/o R; according to the experimental performances, the impact of PRIM w/o R is prominently important because approximately three quarters experiments achieve the suboptimal results second to our PRIM model. This is due to that the multi-hop propagation probability estimation truly simulates the process of information dissemination and is helpful to filter out the nodes with the higher transmission ability.

Role Structure Embedding Role structure embedding behaves temperamentally, as is shown in PRIM w/o P in Table 3. Sometimes it achieves the same best results as our PRIM model in Cora ML dataset and second best performance in PPI dataset. For Jazz data it seems acting not so good and this may originates from the high average degree distribution and smallest network size and this results in the low distinction of role features.

Table 3. Results (diffusion count %) of different variants for PRIM on some datasets under weighted IC model. The best is marked in **bold blue** while the second best is in underline.

Model \ Dataset	Cora ML				Jazz				PPI			
	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
w/o R	11.9	27.6	39.4	52.9	9.6	27.7	38.7	53.0	46.3	56.0	64.9	**75.1**
w/o P	13.2	**28.5**	38.7	**53.1**	2.9	28.7	38.1	52.5	46.5	56.4	64.2	74.8
PRIM	**15.7**	**28.5**	**39.9**	**53.1**	**14.7**	**29.7**	**41.9**	**55.2**	**47.8**	**57.1**	**65.0**	74.6

Stability Analysis The spreading process is random under the weighted IC model. To evaluate the stability of our model, we resample different sizes of test data and figure out the optimal seed nodes to calculate the diffusion account. From Fig. 2, the performance of each sizes of seed nodes shows no difference, which proves the effectiveness of our model. Besides, the expected influence spread increases with seed number grows, which is consistent with the monotone increasing characteristic of influence spread under the weighted IC model.



(a) Jazz (b) Cora ML (c) PPI

Fig. 2. The diffusion account of the datasets for different sample sizes.

5 Conclusion

This paper proposes a novel framework, PRIM built with a feature extractor module and a multi-head message passing model, to resolve the IM problem. PRIM has obtained satisfactory performances from the perspective of global propagation characteristics modeling and role-oriented structural embedding. The propagation probability embedding module well modeled the underlying information propagation process and the role structure embedding distinguished the dissemination capacity of influential and non-influential nodes. The multi-head attention message passing module well learned the nodes' characteristics in the spreading process and greatly selected the optimal seed nodes in a more efficient way. Our extensive experiments on some real datasets demonstrate our PRIM's higher efficacy and efficiency over the state-of-the-art IM methods.

Acknowledgments. This work was partially supported by National Key R&D Program of China (No. 2022YFB3102600), NSFC (Nos. 62272469, 62306322, 72301284), and The Science and Technology Innovation Program of Hunan Province (No. 2023RC1007).

References

1. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. nature **406**(6794), 378–382 (2000)
2. Azaouzi, M., Mnasri, W., Romdhane, L.B.: New trends in influence maximization models. Computer Science Review **40**, 100393 (2021)
3. Bavelas, A.: Communication patterns in task-oriented groups. The journal of the acoustical society of America **22**(6), 725–730 (1950)
4. Boldi, P., Vigna, S.: Axioms for centrality. Internet Mathematics **10**(3-4), 222–262 (2014)
5. Bonacich, P.: Power and centrality: A family of measures. American journal of sociology **92**(5), 1170–1182 (1987)
6. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., et al.: Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic acids research **31**(9), 2443–2450 (2003)
7. Chen, T., Liu, W., Fang, Q., Guo, J., Du, D.Z.: Minimizing misinformation profit in social networks. IEEE Transactions on Computational Social Systems **6**(6), 1206–1218 (2019)
8. Chen, T., Yan, S., Guo, J., Wu, W.: Touplegdd: A fine-designed solution of influence maximization by deep reinforcement learning. IEEE Transactions on Computational Social Systems **11**(2), 2210–2221 (2023)
9. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proc. of SIGKDD. pp. 1029–1038 (2010)
10. Cheng, L., Zhu, P., Tang, K., Gao, C., Wang, Z.: Gin-sd: source detection in graphs with incomplete nodes via positional encoding and attentive fusion. In: Proc. of AAAI. vol. 38, pp. 55–63 (2024)
11. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proc. of SIGKDD. pp. 57–66 (2001)
12. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry (1977)
13. Gasteiger, J., Weißenberger, S., Günnemann, S.: Diffusion improves graph learning. Advances in neural information processing systems **32** (2019)
14. Guo, Q., Wang, S., Wei, Z., Chen, M.: Influence maximization revisited: Efficient reverse reachable set generation with bound tightened. In: Proc. of SIGMOD. pp. 2167–2181 (2020)
15. Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., Li, L.: Rolx: structural role extraction & mining in large graphs. In: Proc. of SIGKDD. pp. 1231–1239 (2012)
16. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proc. of SIGKDD. pp. 137–146 (2003)
17. Kumar, S., Mallik, A., Khetarpal, A., Panda, B.S.: Influence maximization in social networks using graph embedding and graph neural network. Information Sciences **607**, 1617–1636 (2022)

18. Kumar, S., Mallik, A., Panda, B.: Influence maximization in social networks using transfer learning via graph-based lstm. Expert Systems with Applications **212**, 118770 (2023)
19. Li, H., Xu, M., Bhowmick, S.S., Rayhan, J.S., Sun, C., Cui, J.: Piano: Influence maximization meets deep reinforcement learning. IEEE Transactions on Computational Social Systems **10**(3), 1288–1300 (2022)
20. Li, Y., Gao, H., Gao, Y., Guo, J., Wu, W.: A survey on influence maximization: From an ml-based combinatorial optimization. ACM Transactions on Knowledge Discovery from Data **17**(9), 1–50 (2023)
21. Li, Y., Fan, J., Wang, Y., Tan, K.L.: Influence maximization on social graphs: A survey. IEEE Transactions on Knowledge and Data Engineering **30**(10), 1852–1872 (2018)
22. Ling, C., Jiang, J., Wang, J., Thai, M.T., Xue, R., Song, J., Qiu, M., Zhao, L.: Deep graph representation learning and optimization for influence maximization. In: Proc. of ICML. pp. 21350–21361. PMLR (2023)
23. McCallum, A.K., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. Information Retrieval **3**, 127–163 (2000)
24. Morone, F., Makse, H.A.: Influence maximization in complex networks through optimal percolation. Nature **524**(7563), 65–68 (2015)
25. Panagopoulos, G., Malliaros, F.D., Vazirgiannis, M.: Multi-task learning for influence estimation and maximization. IEEE Transactions on Knowledge and Data Engineering **34**(9), 4398–4409 (2020)
26. Rossi, R., Ahmed, N.: The network data repository with interactive graph analytics and visualization. In: Proc. of AAAI. vol. 29 (2015)
27. Sumith, N., Annappa, B., Bhattacharya, S.: Influence maximization in large social networks: Heuristics, models and parameters. Future Generation Computer Systems **89**, 777–790 (2018)
28. Tang, J., Tang, X., Xiao, X., Yuan, J.: Online processing algorithms for influence maximization. In: Proc. of SIGMOD. pp. 991–1005 (2018)
29. Wang, C., Zhao, J., Li, L., Jiao, L., Liu, J., Wu, K.: A multi-transformation evolutionary framework for influence maximization in social networks. IEEE Computational Intelligence Magazine **18**(1), 52–67 (2023)
30. Wang, J., Jiang, J., Zhao, L.: An invertible graph diffusion neural network for source localization. In: Proc. of WWW. pp. 1058–1069 (2022)
31. Xia, W., Li, Y., Wu, J., Li, S.: Deepis: Susceptibility estimation on social networks. In: Proc. of WSDM. pp. 761–769 (2021)
32. Ye, M., Liu, X., Lee, W.C.: Exploring social influence for recommendation: a generative model approach. In: Proc. of SIGIR. pp. 671–680 (2012)
33. Zhang, J., Wang, B.: Encoding node diffusion competence and role significance for network dismantling. In: Proc. of WWW. pp. 111–121 (2023)
34. Zhou, B., Pei, S., Muchnik, L., Meng, X., Xu, X., Sela, A., Havlin, S., Stanley, H.E.: Realistic modelling of information spread using peer-to-peer diffusion patterns. Nature Human Behaviour **4**(11), 1198–1207 (2020)