A Probabilistic Graph Diffusion Model for Source Localization (Student Abstract)

Tangjiang Qian[1*], **Xovee Xu**[1*], **Zhe Xiao**[2], **Ting Zhong**[1,3†], **Fan Zhou**[1,2]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China,
[2]Science and Technology on Communication Networks Laboratory, Shijiazhuang, Hebei 050050, China,
[3]Kashi Institute of Electronics and Information Industry, Kashi, Xinjiang 844000, China
tj.qian@std.uestc.edu.cn, xovee@ieee.org, xiaozhe5401@126.com, {zhongting, fan.zhou}@uestc.edu.cn

Abstract

Source localization, as a reverse problem of graph diffusion, is important for many applications such as rumor tracking, detecting computer viruses, and finding epidemic spreaders. However, it is still under-explored due to the inherent uncertainty of the diffusion process: after a long period of propagation, the same diffusion process may start with diverse sources. Most existing solutions utilize deterministic models and therefore cannot describe the diffusion uncertainty of sources. Moreover, current probabilistic approaches are hard to conduct smooth transformations with variational inference. To overcome the limitations, we propose a probabilistic framework using continuous normalizing flows with invertible transformations and graph neural networks to explicitly model the uncertainty of the diffusion source. Experimental results on two real-world datasets demonstrate the effectiveness of our model over strong baselines.

Introduction

Graph diffusion prediction is an important task in social networks and graph mining, which aims to unveil the propagation patterns of information and predict its future state. On the contrary, source localization is a reverse problem of graph diffusion and tries to identify the source(s) of the observed diffusion process. Source localization plays a key role in many practical situations, such as misinformation/rumor detection in social networks, epidemic control of infectious diseases, and isolated failures in smart grids.

Although prior studies have made significant improvements on source localization, they still face several challenges. First, most of the existing methods focus on deterministic learning to solve the problem, which are unable to handle the diffusion uncertainty of the source. For example, different diffusion sources can generate the same diffusion observations after a long time interval, making the diffusion pattern matching a hard problem for the model to solve. Second, current probabilistic methods are inadequate to conduct smooth transformations between latent space and data distribution with variational inference models. When the diffusion graph grows large and becomes complex, transforming the data distribution to multivariate Gaussian is challenging.

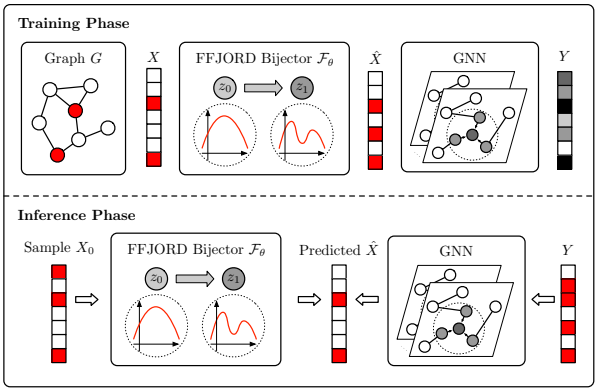

Figure 1: The proposed method consists of two phases. The training phase constructs the graph diffusion model, and the inference phase infers sources from diffused observations.

In this work, we propose a probabilistic graph diffusion method for source localization, which tackles the uncertainty problem and learns the diffusion patterns using deep generative model and graph neural networks (GNNs). The overall framework of our method is depicted in Figure 1.

Methodology

Problem Definition Give an undirected graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. Let $Y = \{y_i\}_i \in \mathbb{R}^{|V|}$ be the infection state vector. $y_i = 1$ if the node i is infected, and $y_i = 0$ otherwise. Let $X = \{x_i\} \in \mathbb{R}^{|V|}$ be the diffusion source vector, $x_i = 1$ if the node i is the source node and $x_i = 0$ otherwise. Due to the highly uncertain process of the diffusion, we build a probabilistic model $p(X|Y)$ to explicitly measure the uncertainty. Since the diffused observation Y depends on the graph topology G, we have the conditional probability $p(X|Y, G)$ according to Bayes rules. The graph diffusion source localization problem can be defined as a Maximum A Posterior (MAP) estimation problem:

$$\hat{X} = \max_X p(X|Y, G) = \max_X p(Y|X, G)p(X) \quad (1)$$

where $\hat{X}$ denotes the predicted source vector.

*These authors contributed equally.
†Corresponding author.

Training Phase Estimating the distribution of diffusion source nodes $p(X)$ is difficult according to Eq. (1). Hence, we leverage a deep generative model (Kingma, Welling et al. 2019) to map the high-dimensional $p(X)$ into low-dimensional $p(Z)$, where $Z \in \mathbb{R}^d$ is the latent random variable vector. The posterior $p(Z|X, Y, G)$ can be used to infer the latent variable Z, but $p(X)$ is intractable. We instead approximate the posterior $q_\theta(Z|X, Y, G)$ parameterized by θ, i.e., we compute the KL-divergence between $p(Z|X, Y, G)$ and $q_\theta(Z|X, Y, G)$. Since in most of the graph diffusion cases the latent variable Z is independent of the diffused observation Y (Ling et al. 2022), the posterior $q(Z|X, Y, G)$ and likelihood $p(X, Y, G|Z)$ can be simplified as $q(Z|X)$ and $p(Y|X, G)p(X|Z)$, respectively.

We then exploit an invertible neural network called FFJORD bijector $\mathcal{F}_\theta$ (Grathwohl et al. 2019) in the generative model to parameterize the distribution $p(X)$, which conducts a series of smooth and invertible transformations between the latent Z and target X. The bijector $\mathcal{F}_\theta$ considers a continuous transformation from latent state $z(t_0)$ to $z(t_1)$ as follows:

$$\log p(z(t_1)) = \log p(z(t_0)) - \int_{t_0}^{t_1} \mathrm{Tr}\left(\frac{\partial \mathcal{F}_\theta}{\partial z(t)}\right) dt. \quad (2)$$

The integration can be solved by ordinary differential equations. The objective function is:

$$\begin{aligned}\mathcal{L}_{\text{train}} = \min_{\theta,\phi} \{ & -\mathbb{E}_{q_\theta}\left[\log p_\phi(Y|X, G) + \log p_\theta(X|Z)\right] \\ & + \mathbb{D}_{\text{KL}}\left[q_\theta(Z|X)||p(Z)\right] \\ & - \mathbb{E}_{q_\theta}\left[\log p(z(t_0)) - \log p(z(t_1))\right] \}.\end{aligned} \quad (3)$$

In practice, the posterior $q_\theta(Z|X)$ and the likelihood $p_\theta(X|Z)$ is parameterized by the FFJORD bijector $\mathcal{F}_\theta$ and $p_\phi(Y|X, G)$ is modeled by the GNN.

Inference Phase Since the distribution $p(X)$ is modeled by $p(Z)$ after training, we can solve the MAP in Eq. (1) via $p(X) = p(X|Z)p(Z)$. However, due to the computational complexity of sampling Z from $p(Z)$, we propose to sample Z from the posterior $q(Z|X)$. The objective function of the inference phase is defined as:

$$\begin{aligned}\mathcal{L}_{\text{infer}} = \min_X \{ & -\log p_\phi(Y|X, G) \\ & - \log[p_\theta(X|Z)q_\theta(Z|\tilde{X})]\},\end{aligned} \quad (4)$$

where $\tilde{X}$ is the source vector from the training data. In practice, we sample an initial diffusion source X_0 from a binomial distribution in which the probability τ is set to 0.5. Then we optimize the value of X following Eq. (4).

Experiments

We conducted the experiments on two real-world source localization datasets Cora-ML and Power-Grid. We randomly select 10% of nodes as the sources and simulate the information diffusion process based on susceptible-infected (SI) and susceptible-infected-recovery (SIR) algorithms.

We use three strong source localization models as the baselines, including LPSI (Wang et al. 2017), GCNSI (Dong et al. 2019), and SL-VAE (Ling et al. 2022). Following previous studies, we use Recall and F1-Score as the evaluation metrics, as source localization is in essential an unbalanced classification problem that needs to retreive the source node from many other nodes.

Dataset	Method	SI		SIR	
		Recall	F1	Recall	F1
Cora-ML	LPSI	0.595	0.247	0.478	0.175
	GCNSI	0.362	0.178	0.338	0.173
	SL-VAE	0.899	0.697	0.562	0.611
	Ours	**0.949**	**0.725**	**0.950**	**0.726**
Power-Grid	LPSI	0.495	0.474	0.472	0.478
	GCNSI	0.348	0.210	0.237	0.153
	SL-VAE	0.932	0.721	0.646	0.665
	Ours	**0.963**	**0.731**	**0.944**	**0.734**

Table 1: Performance comparison of our model and baselines on two datasets under SI and SIR diffusion algorithms.

We evaluate the performance of our model and compare it with other source localization approaches under both SI and SIR diffusion algorithms, the results are shown in Table 1. We can see that our model significantly outperforms LPSI and GCNSI and achieves non-trivial improvements against SL-VAE in terms of both metrics. Besides, all three baselines are worse in SIR than in SI, because the diffusion process of SIR is more complex than SI's. These results verify our motivation of handling the uncertainties in graph diffusion and source localization by constructing a probabilistic model using continuous normalizing flows.

Acknowledgments

This work was supported in part by NSF of Sichuan Province (Grant No. 2022NSFSC0505), National Natural Science Foundation of China (Grant No. 62176043 and No. 62072077), and Foundation of Science and Technology on Communication Networks Laboratory (Grant No. FWX22641X001).

References

Dong, M.; Zheng, B.; Hung, N. Q. V.; Su, H.; and Li, G. 2019. Multiple Rumor Source Detection with Graph Convolutional Networks. In *CIKM*, 569–578.

Grathwohl, W.; Chen, R.; Bettencourt, J.; Sutskever, I.; and Duvenaud, D. 2019. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *ICLR*.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Ling, C.; Liang, J.; Wang, J.; and Zhao, L. 2022. Source Localization of Graph Diffusion via Variational Autoencoders for Graph Inverse Problems. In *SIGKDD*, 1010–1020.

Wang, Z.; Wang, C.; Pei, J.; and Ye, X. 2017. Multiple Source Detection without Knowing the Underlying Propagation Model. In *AAAI*, 217–223.