A Generalized Diffusion Framework with Learnable Propagation Dynamics for Source Localization

Dongpeng Hou[1,2], **Yuchen Wang**[2], **Chao Gao**[2] and **Xianghua Li**[2*]
[1]School of Mechanical Engineering, Northwestern Polytechnical University
[2]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
li_xianghua@nwpu.edu.cn

Abstract

Source localization has been widely studied in recent years due to its crucial role in controlling the spread of harmful information. Existing methods only achieve satisfactory performance within a specific propagation model, which restricts their applicability and generalizability across different scenarios. To address this, we propose a Generalized Diffusion Framework for Source Localization (GDFSL), which enhances probabilistic diffusion models to flexibly capture the underlying dynamics of various propagation scenarios. By redefining the forward diffusion process, GDFSL ensures convergence to a real distribution of infected states that accurately represents the targeted dynamics, enabling the model to learn unbiased noise in a self-supervised manner that encodes fine-grained propagation characteristics. A closed-form reverse diffusion process is then derived to trace the propagation back to the source. The process does not rely on an explicit source label term, facilitating direct inference of sources from observed data. Experimental results show that GDFSL outperforms SOTA methods in various propagation models, particularly in scenarios where historical training data is limited or unavailable. The code is available at https://github.com/cgao-comp/GDFSL.

1 Introduction

In today's information age, network science is ubiquitous across various domains, including online social media and offline human societal systems [Ling *et al.*, 2022; Zhu *et al.*, 2024]. The harmful propagation processes within these complex systems, such as misinformation diffusion, widespread viral spread, and large-scale cascading failures, profoundly impact societal stability, economic efficiency, and security [Zhu *et al.*, 2022]. Identifying the propagation source and controlling the propagation chain is an effective measure [Yan *et al.*, 2024]. Consequently, the source localization study becomes crucial for managing and mitigating these issues.

Snapshot based source localization is widely focused due to the convenience of snapshot capture [Jiang *et al.*, 2016; Hou *et al.*, 2023]. Based on Bayesian theory, the localization problem can be demonstrated as the Maximum Posterior Estimate (MPE) or the Maximum Likelihood Estimate (MLE) [Huang *et al.*, 2018; Shah and Zaman, 2010; Shah and Zaman, 2011]. Among them, the MPE branch focuses on inferring the most likely sources based on observed data, while the MLE branch selects specific sources to reconstruct the observed propagation scenes. Given that the MLE branch involves computationally intensive simulation methods with high complexity, such as the Monte Carlo simulation, current research predominantly concentrates on MPE, including the centrality methods [Wang *et al.*, 2017; Hou *et al.*, 2024a] and deep learning based methods [Ling *et al.*, 2022].

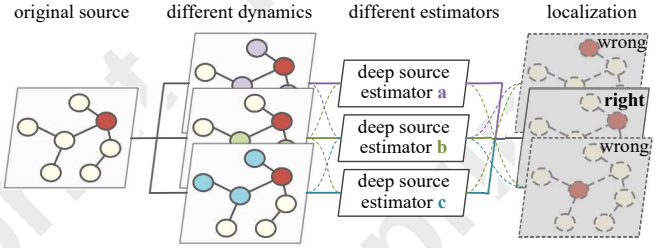

Figure 1: Illustration of source localization under different propagation dynamics. The same source (red node) in various dynamic scenarios can lead to different spreading cascades. For each estimator, it can accurately infer the source (right) in some propagation scenarios, while in others, it fails to infer the sources (wrong), highlighting the challenges in source localization across diverse dynamics.

Many existing source localization methods heavily depend on the assumption of specific propagation models, which specifies their applicability to certain propagation dynamics and limits their effectiveness and transferability in other scenarios. [Zhu and Ying, 2014; Zhu *et al.*, 2017]. Some works do not need to rely on underlying propagation dynamics for source localization [Dong *et al.*, 2019; Wang *et al.*, 2022; Hou *et al.*, 2025], but many of them still fail to capture the propagation patterns effectively in the source inference process. As a result, the inadequate consideration of the propagation's fine-grained characteristics prevents the model from achieving optimal prediction performance.

Therefore, there is a critical need for a more flexible and generalized framework that can capture various propagation

dynamics and learn different propagation patterns. Diffusion models have proven powerful in solving inverse problems, with the reverse process providing a feasible solution for the inverse propagation (i.e., source localization) [Huang *et al.*, 2023]. And the self-supervised strategy in the forward process incorporates noise modeling, enabling the flexible learning of complex patterns or distributions in the data. Based on this, we develop a Generalized Diffusion Framework for Source Localization (GDFSL) that can use noise to flexibly learn each specific pattern in different propagation scenarios. More specifically, in the forward process, GDFSL redefines the forward process to converge to a fully infected state that accurately reflects the underlying propagation dynamics, and the unbiased noise is further derived to encode the microstates of various propagation mechanisms. Subsequently, the interpretable close-formed reverse diffusion process leverages these learned patterns in the unbiased noise to accurately trace back to the original source. The contributions are as follows:

- We propose a unified localization framework that breaks traditional source localization methods dependent on a specific fixed propagation model assumption. By enabling flexible adaptation to various propagation dynamics, GDFSL expands its applicability and enhances transferability across diverse scenarios.
- GDFSL redefines the forward diffusion process to converge toward a fully infected state that precisely reflects real propagation dynamics. This redefinition enables the derivation of unbiased noise which encodes the microstates of different propagation mechanisms, thus capturing fine-grained characteristics crucial for accurate source inference.
- GDFSL derived an interpretable closed-form reverse diffusion solution to trace the fully infected states back to the original source, which is building on the unbiased noise learned in the forward process. And this closed-form expression does not rely on an explicit source label term, allowing for direct inference of propagation sources from observed data, rather than depending on historical datasets with labeled sources for training as in supervised deep learning methods.

2 Related Work

2.1 Propagation Models

Many propagation models have been proposed to capture the propagation characteristics. These models are simulated on the static social network and generate the propagation datasets for evaluating the performance of localization methods, such as the Susceptible-Infected (SI) model [Yang *et al.*, 2020; Paluch *et al.*, 2021; Zang *et al.*, 2015] and the Susceptible-Infected-Recovered (SIR) model [Zhu and Ying, 2014; Tang *et al.*, 2018]. A variation of SIR is the Susceptible-Infected-Susceptible (SIS) model where recovered individuals can be re-infected [Dong *et al.*, 2019]. However, a key limitation of these traditional models is their homogeneous features. In reality, every individual in social networks has unique features, leading to diverse responses to the same information. Many researchers have further considered the heterogeneous diffusion models that characterize the difference between individuals [Karrer and Newman, 2010; Ellison, 2024]. For instance, models like the Heterogeneous SI (HSI) and Heterogeneous SIR (HSIR) consider varied infection and recovery rates respectively. In addition to these diffusion models, influence models such as the Independent-Cascade (IC) and Linear Threshold (LT) models [Goldenberg *et al.*, 2001; Granovetter, 1978] have been adopted. These models underscore the mutual influence dynamics between individuals. These propagation models attempt to reveal the underlying patterns of information dynamics through different interaction mechanisms and scenario assumptions.

2.2 Source Localization Methods

Due to the convenience and feasibility of snapshot acquisition, many works focus on snapshot based source localization branches [Cheng *et al.*, 2024]. Dong et al. introduce a GCN based source identification model to tackle the multiple rumor source localization problem [Dong *et al.*, 2019]. Additionally, some methods construct the dynamic features of propagation before inferring the source, such as IVGD [Wang *et al.*, 2022], MCGNN [Shu *et al.*, 2021] and SL_VAE [Ling *et al.*, 2022]. Hou et al. employ an encoder-decoder module to learn the user influence matrix [Hou *et al.*, 2023]. Furthermore, considering the powerful performance of the diffusion model in solving inverse problems, Huang et al. propose a denoising diffusion model to quantify the uncertainty in the propagation process to improve the localization performance [Huang *et al.*, 2023]. Yan et al. propose a discrete diffusion model for source localization by designing a reversible residual block with graph convolutional networks [Yan *et al.*, 2024]. Overall, these methods attempt to infer sources without relying on underlying models, which can enhance the application capability in different propagation scenarios. However, not considering the underlying propagation model can easily lead to a lack of deep integration with the propagation dynamics during the source inference process. Developing a generalized source localization model that can effectively learn diverse propagation patterns and achieve accurate localization performance across various scenarios remains a significant challenge.

3 Preliminary

Unlike existing methods that require one or multiple snapshots as input, our method can be applied to scenarios where any number of snapshots for each propagation is available. A series snapshots of each propagation $\{\mathcal{G}^{s_j} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, H^{s_j}) \mid j \geq 1\}$, conveniently available at different timestamps j, are collected, where $\mathcal{V}$ and $\mathcal{E}$ are the node (user) set and edge (relationship) set, respectively. $\mathcal{F}$ is an optional set of node attributes. H denotes the set of node states where $H(v_i) = 1$ if node v_i is observed to be infected or active, and $H(v_i) = 0$ otherwise. And we denote the initial state, viewed from the perspective of only the source being infected or active, as H^{s_0}. Our method first learns the propagation patterns from the observed state H to a fully infected state. Leveraging the learned patterns, an inverse diffusion module traces the fully

infected state back to the original source. And the method aims to predict an initial state $\hat{H}^0$ that maximizes indicators such as $\frac{\{H^{*0}=1\}\cap\{\hat{H}^0=1\}}{\{H^{*0}=1\}\cup\{\hat{H}^0=1\}}$. There are various classical propagation dynamics, and our work aims to capture the underlying patterns from these dynamics to facilitate multi-scenario source localization.

Heterogeneous Diffusion Process (HDP): In HDP, nodes are either informed or uninformed. The source node initiates propagation at time t_s, with only the source informed initially. The message spreads along the shortest paths, with each edge (v_i, v_j) having a propagation delay θ_{v_i,v_j} sampled from a uniform distribution $U(1, h)$, introducing heterogeneity in delays. This models diverse real-world propagation scenarios where different groups experience varying speeds of information spread.

Time-Varying Diffusion Process (TDP): TDP allows the diffusion process to adapt over time. The source can activate at any moment with a probability drawn from $U(0, 1)$. Once an edge is activated, it incurs a delay $\theta_{v_i,v_j} \sim U(0, h)$ and becomes invalid after θ_{v_i,v_j} time units, preventing further propagation along that edge. This reflects real-world interactions where time constraints limit message dissemination.

Random Walk Propagation Process (RDP): In RDP, the message performs a random walk in the network. When the message arrives at node v at time t, it jumps to a randomly selected neighbor v_j with probability $1/k_v$, where k_v is the degree of v. The message resides on one node at a time, simulating scenarios such as sequential conversations.

Linear Threshold (LT) Model: The LT model is a classical diffusion model where each node becomes informed if the cumulative influence from its informed neighbors exceeds a certain threshold.

Susceptible-Infected (SI) Model: Each node is either susceptible or infected stat. At each time step, an infected node transmits the message to each susceptible neighbor with a probability.

Independent Cascade (IC) Model: Once a node becomes infected, it has a single opportunity to infect each of its susceptible neighbors.

4 Method

In this section, we introduce the Generalized Diffusion Framework for Source Localization (GDFSL). Initially, we re-derive the forward diffusion process under the condition that the Gaussian noise distribution has a mean of one. This adaptation ensures that the forward diffusion process accurately reflects the underlying dynamics of information over a complete propagation cycle, from no infection to full infection within the network. We then derive the reverse diffusion process of conditional probabilities, providing a robust theoretical foundation for GDFSL.

4.1 Forward Process of GDFSL

The forward process of the diffusion model is mathematically described by the following equations [Ho *et al.*, 2020]:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right), \tag{1}$$

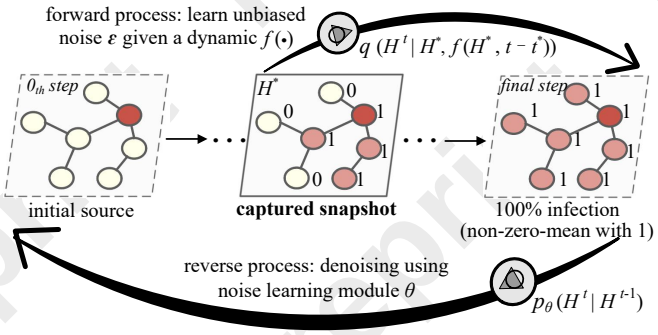

Figure 2: Interpretability of the GDFSL in source detection tasks. The forward process starts from t^* to T, learning the changes of node states during the corresponding time step under the propagation dynamics. The reverse process directly infers the propagation source from the observation using a closed-form solution back to t_0.

$$q\left(x_{1:T} \mid x_0\right)=\prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right). \tag{2}$$

These equations define the probabilistic progression from any state x_{t-1} to x_t, where β_t are variance parameters that control the noise level at each step, progressively adding uncertainty as the process evolves. Generally, $\beta_1 << \beta_T$.

To enhance the applicability of GDFSL to the source localization task, we introduce $\vec{\mathbf{1}}$ as the vector of all ones, whose dimensionality matches that of the state vectors x_t and x_{t-1}. Then, we use $\vec{\mathbf{1}}$ as guiding component to adjust the latent distribution space:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}+(1-\sqrt{1-\beta_t})\vec{\mathbf{1}}, \beta_t \mathbf{I}\right), \tag{3}$$

$$q\left(x_{1:T} \mid x_0\right)=\prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right). \tag{4}$$

The guiding component $\vec{\mathbf{1}}$ as a fully infected knowledge ensures the ability of the reverse process to trace the original source.

Theorem 1. $T \to \infty, x_T \sim \mathcal{N}(\vec{\mathbf{1}}, \mathbf{I})$ in Eq. (3).

Proof: Without loss of generality, we define $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Then starting from the iterative diffusion equation:

$$\begin{aligned}\mathbf{x}_t =& \sqrt{\alpha_t}\mathbf{x}_{t-1} + (1-\sqrt{\alpha_t})\vec{\mathbf{1}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\ =& \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + (1-\sqrt{\alpha_t}\sqrt{\alpha_{t-1}})\vec{\mathbf{1}} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \\ =& \cdots \\ =& \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1-\sqrt{\bar{\alpha}_t})\vec{\mathbf{1}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\end{aligned}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ represents the product of the scaling factors up to time t, $\boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\bar{\boldsymbol{\epsilon}}_{t-2}$ merges two Gaussians.

As $T \to \infty$, the product $\sqrt{\bar{\alpha}_T}$ tends to zero due to the properties of the parameters α_t which are designed such that $0 \le \alpha_t < 1$. This leads to $\sqrt{\bar{\alpha}_T}\mathbf{x}_0$ vanishing, and the expression simplifies to:

$$\lim_{T\to\infty} \mathbf{x}_T = (1-\sqrt{\bar{\alpha}_T})\vec{\mathbf{1}} + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}.$$

Since $(1-\sqrt{\bar{\alpha}_T}) \to 1$ as $T \to \infty$, $\mathbf{x}_T$ converges to $\vec{\mathbf{1}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, $x_T \sim \mathcal{N}(\vec{\mathbf{1}}, \mathbf{I})$ as $T \to \infty$. ■

Then we can get the closed-form solution of the optimized forward process for the input of the source localization task as follows:

$$q\left(H^t \mid H^0\right) = \mathcal{N}\left(H^t; \sqrt{\bar{\alpha}_t} H^0 + (1-\sqrt{\bar{\alpha}_t})\vec{\mathbf{1}}, (1-\bar{\alpha}_t)\mathbf{I}\right). \tag{5}$$

where H represents the state set of user participation, indicating whether users are active or inactive. H^0 denotes the node state set at the initial step, where only the earliest source user is active, and H^t denotes the noised state at the t_{th} step in the diffusion process.

It is essential to determine the step index $t^* \in (t_0, T)$ of the snapshot within the forward propagation sequence when a snapshot $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, H^*)$ is given. We can easily get the step index of a given uniform propagation input $\mathcal{G}$:

$$t^* = \frac{\sum_{H^*} I(H^*(v) = 1)}{|\mathcal{V}|} * T, \tag{6}$$

where $\sum_{H^*} I(H^*(v) = 1)$ represents the sum of the indicator function I, which equals 1 if the condition $H^*(v) = 1$ for user v is met.

Then we can get the closed-form conditional forward process after $(t - t^*)$ steps by adding noise to the given user state H^*.

$$\begin{aligned} q\left(H^t \mid H^*\right) = \mathcal{N}\Big(H^t; \\ &\sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{t^*}} H^* + (1 - \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{t^*}})\, \vec{\mathbf{1}}, \\ &(1 - \alpha_t \alpha_{t-1} \cdots \alpha_{t^*})\, \mathbf{I}\Big), \ \text{ s.t. } t^* < t \leq T, \\ = \mathcal{N}\Big(H^t; &\sqrt{\bar{\alpha}_{tt^*}} H^* + \left(1 - \sqrt{\bar{\alpha}_{tt^*}}\right) \vec{\mathbf{1}}, \\ &(1 - \bar{\alpha}_{tt^*})\, \mathbf{I}\Big), \ \text{ s.t. } t^* < t \leq T. \end{aligned} \tag{7}$$

where $\bar{\alpha}_{tt^*} = \alpha_t \alpha_{t-1} \cdots \alpha_{t^*}$. Based on Eq. (7), after giving a state set H^* from snapshot $\mathcal{G}$ and a random subsequent step $t > t^*$, the forward propagation process can be completed in an interval of $(t - t^*)$ steps.

Based on the closed-form solution of the forward process, we can use noise to store and encode the microstates of different propagation dynamics. Instead of randomly initializing the noise, which is widely acknowledged to introduce bias, we assign the noise based on the state transitions in the corresponding propagation dynamics. Specifically, we leverage a propagation mechanism $f(\cdot)$ to transform the observed state H^* at t^* into an unbiased state H^t_{real} at t in the scenarios of the corresponding dynamics. Formally,

$$H^t_{\text{real}} = f\big(H^*, t - t^*\big), \tag{8}$$

where $f(\cdot)$ reflects the particular propagation dynamics under consideration (e.g., SI, IC, LT, etc.).

After obtaining H^t_{real}, we substitute it into Eq. (7) to solve for the unbiased noise term $\boldsymbol{\epsilon}$.

$$\boldsymbol{\epsilon} = \frac{H^t_{\text{real}} - \sqrt{\bar{\alpha}_{tt^*}}\, H^* - \left(1 - \sqrt{\bar{\alpha}_{tt^*}}\right) \vec{\mathbf{1}}}{\sqrt{1 - \bar{\alpha}_{tt^*}}}, \quad t^* < t \leq T. \tag{9}$$

In this way, the noise $\boldsymbol{\epsilon}$ is no longer an arbitrary random variable, instead, the state transitions from uninfected to infected are preserved and reflected in the noise distribution from H^* to H^t. By learning $\boldsymbol{\epsilon}$ in such a forward process, the model can effectively capture and distinguish the various state transitions under different propagation mechanisms.

4.2 Reverse Process of GDFSL

As defined in Sec. 4.1, the optimized noise addition process enables the forward process to effectively realize the propagation dynamics. Then, the conditional probabilities for the reverse diffusion process need to be considered. We need to derive the revised $q(H^{t-1}|H^t)$.

Consistent with the traditional diffusion model [Ho *et al.*, 2020], the conditional probability distribution of the reverse process is as follows:

$$\begin{aligned} q(H^{t-1}|H^t, H^0) &= \frac{q(H^{t-1}, H^t, H^0)}{q(H^t, H^0)} \\ &= \frac{q(H^0) q(H^{t-1}|H^0) q(H^t|H^{t-1}, H^0)}{q(H^0) q(H^t|H^0)} \\ &= q(H^t|H^{t-1}, H^0) \frac{q(H^{t-1}|H^0)}{q(H^t|H^0)}. \end{aligned} \tag{10}$$

Given that the forward process is a Markov chain, $q(H^t|H^{t-1}, H^0)$ is independent of H^0, thus we can get:

$$q(H^{t-1}|H^t, H^0) = q(H^t|H^{t-1}) \frac{q(H^{t-1}|H^0)}{q(H^t|H^0)}. \tag{11}$$

In Eq. (11), $q\left(H^t \mid H^0\right)$ is shown in Eq. (5), we can also get the other two forward equations as follows:

$$q(H^t|H^{t-1}) = \mathcal{N}(H^t; \sqrt{\alpha_t} H^{t-1} + (1-\sqrt{\alpha_t})\vec{\mathbf{1}}, \beta_t \mathbf{I}). \tag{12}$$

$$q(H^{t-1}|H^0) = \mathcal{N}(H^{t-1}; \sqrt{\bar{\alpha}_{t-1}} H^0 + (1-\sqrt{\bar{\alpha}_{t-1}})\vec{\mathbf{1}}, (1-\bar{\alpha}_{t-1})\mathbf{I}). \tag{13}$$

Based on the Gaussian probability density function, Eq. (11) can be further expanded as follows:

$$\begin{aligned} &q(H^{t-1}|H^t, H^0) \propto \exp \\ &\Bigg\{-\frac{1}{2}\Bigg(\Big(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\Big)(H^{t-1})^2 - \Bigg(\frac{2\sqrt{\alpha_t}}{\beta_t}(H^t - (1-\sqrt{\alpha_t})\vec{\mathbf{1}}) \\ &+ \frac{2}{1-\bar{\alpha}_{t-1}}(\sqrt{\bar{\alpha}_{t-1}} H^0 + (1-\sqrt{\bar{\alpha}_{t-1}})\vec{\mathbf{1}})\Bigg) H^{t-1} + C(H^t, H^0)\Bigg)\Bigg\}. \end{aligned} \tag{14}$$

Given the properties of the Gaussian distribution, the variance can be determined as follows:

$$\sigma^2_{t-1|t} = \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)^{-1} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t. \tag{15}$$

And the mean can be determined as follows:

$$\begin{aligned} \mu_{t-1|t} = \frac{\sigma^2_{t-1|t}}{2} \cdot \Bigg(\frac{2\sqrt{\alpha_t}}{\beta_t}(H^t - (1-\sqrt{\alpha_t})\vec{\mathbf{1}}) + \\ \frac{2}{1-\bar{\alpha}_{t-1}}(\sqrt{\bar{\alpha}_{t-1}} H^0 + (1-\sqrt{\bar{\alpha}_{t-1}})\vec{\mathbf{1}})\Bigg) = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} H^t \\ + \frac{\beta_t \sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t} H^0 + \left(\frac{(1-\sqrt{\bar{\alpha}_{t-1}})(1-\sqrt{\alpha_t})(1-\sqrt{\bar{\alpha}_t})}{1-\bar{\alpha}_t}\right)\vec{\mathbf{1}}. \end{aligned} \tag{16}$$

Combining with Eq. (5), Eq. (16) for $\mu_{t-1|t}$ can be further demonstrated as follows:

$$\mu_{t-1|t} = \frac{1}{\sqrt{\alpha_t}}\left(H^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right) - \left(\frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}(1-\sqrt{\bar{\alpha}_t})}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)} - \frac{(1-\sqrt{\bar{\alpha}_{t-1}})(1-\sqrt{\alpha_t})(1-\sqrt{\bar{\alpha}_t})}{1-\bar{\alpha}_t}\right)\vec{\mathbf{1}}. \tag{17}$$

Thus, the variance $\sigma^2_{t-1|t}$ of the posterior conditional Gaussian distribution is given by Eq. (15), and the mean $\mu_{t-1|t}$ is expressed by Eq. (17). With these closed-form solutions established, we can train GDFSL for source inference, as demonstrated in Alg. 1. Training proceeds as follows: Lines **2-11** of Alg. 1 detail the loss computation for unbiased noise learning, specifically tailored to the given propagation pattern f. This ensures that the learned propagation dynamics are consistent with the theoretical foundation provided by our derivation based on the non-zero-mean Gaussian model. Subsequently, lines **12-19** incorporate additional observational constraints by aligning the node states generated from the reverse diffusion process with the corresponding steps and observed snapshots.

It's worth mentioning that for propagation if there are more than two captured snapshots at different timestamps, lines **2-11** of Alg. 1 can be executed based on multiple different t^*. Subsequently, during the reverse generation process, alignment between the generative and observed snapshot can be made multiple times at the corresponding t^* step. Therefore, GDFSL can leverage and integrate any arbitrary number of multiple observations from a propagation, such flexibility enhances the model's robustness across different observed timestamps of different propagations.

Furthermore, the reverse diffusion process from T to t_1 can realize source prediction. After T iterations, H^0 represents the predicted initial state of all users involved in the given propagation. The node exhibiting the highest probability in H^0 is identified as the predicted source. The source inference process is detailed in Alg. 2.

Algorithm 1 Training for GDFSL

Input: Initial parameters: α_t, β_t for $t = 1, \ldots, T$; A snapshots collection of K propagation cascades $\{\{\mathcal{G}^{(i,s_j)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{F}^{(i)}, H^{(i,s_j)})\}_{j\geq 1}\}_{i=1}^{K}$.
Output: Optimized parameters θ of H^0 prediction module.

1: **repeat** for each of the K propagation cascades, i.e., $\mathcal{G}^{(i)} = \{\mathcal{G}^{(i,s_j)}{=}(\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{F}^{(i)}, H^{(i,s_j)})\}_{j\geq 1}$
2: **for** $(\mathcal{V}, \mathcal{E}, \mathcal{F}, H^*) \sim q\left(\mathcal{G}^{(i)}\right)$ **do**
3: $\mathcal{F}^* \leftarrow (\mathcal{F}, H^*)$
4: Determine the step index of H^*, i.e., $t^* = \frac{\sum_{H^*} I(H^*(v)=1)}{|\mathcal{V}|} * T$ //Eq. (6)
5: $t \sim$ Uniform$\{t^*, \ldots, T\}$
6: Compute the time interval $tt^* = t - t^*$
7: Execute the given propagation patterns $H^t_{\text{real}} = f\left(H^*, t - t^*\right)$ //Eq. (8)
8: Using the unbiased noise to record the propagation mechanism $\boldsymbol{\epsilon}{=}\frac{H^t_{\text{real}} - \sqrt{\bar{\alpha}_{tt^*}}\, H^* - \left(1-\sqrt{\bar{\alpha}_{tt^*}}\right)\vec{\mathbf{1}}}{\sqrt{1-\bar{\alpha}_{tt^*}}}$ //Eq. (9)
9: Define a denoising module $\epsilon_\theta(H^t_{\text{real}}, tt^*, \mathcal{V}, \mathcal{E}, \mathcal{F}^*)$
10: Take gradient descent step on: $\nabla_\theta \left\|\epsilon - \epsilon_\theta\right\|^2$
11: **end for**
12: $H^T \sim \mathcal{N}(\vec{\mathbf{1}}, \mathbf{I}), \hat{H}^T \leftarrow H^T$
13: **for** $t = T, \ldots, 1$ **do**
14: $z \sim \mathcal{N}(0, \mathbf{I})$
15: $\sigma^2_{t-1|t} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$ //Eq. (15)
16: $\mu_{t-1|t} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{H}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{H}^t, t, \mathcal{V}, \mathcal{E}, \mathcal{F}^*)\right) - \left(\frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}(1-\sqrt{\bar{\alpha}_t})}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)} \frac{(1-\sqrt{\bar{\alpha}_{t-1}})(1-\sqrt{\alpha_t})(1-\sqrt{\bar{\alpha}_t})}{1-\bar{\alpha}_t}\right)\vec{\mathbf{1}}$ //Eq. (17)
17: $\hat{H}^{t-1} = \mu_{t-1|t} + \sigma_{t-1|t} \cdot z$
18: **end for**
19: Take gradient descent step on: $\nabla_\theta \left\|H^* - \hat{H}^*\right\|^2$ for each observed snapshot at t^*
20: **until** converged
21: **return** optimized parameters θ

5 Experiments

5.1 Datasets and Setting

We use three datasets collected from two real-world social media platforms, Weibo and Twitter, for source localization tasks, namely Weibo [Ma *et al.*, 2017], Twitter15, and Twitter16 [Liu *et al.*, 2015; Ma *et al.*, 2016]. F in the dataset includes user description, blue verification status, location, registration date, number of posts, fans list, and followings list [Hou *et al.*, 2024b]. The user scale ranges from a minimum of 10 to a maximum of 57,186, with an average user scale of 883 across all cascades. Additionally, the cascades vary in propagation depth, ranging from a minimum of 2 to a maximum of 22, with an average depth of 6.15. The relevant information of the three datasets is shown in Tab. 1. And to demonstrate the generalizability of our proposed method, we employ widely recognized simulated synthetic datasets for source localization, generated based on the HDP, TDP, RDP, LT, SI, and IC propagation models, as commonly used in traditional source localization methods [Ling *et al.*, 2022; Wang *et al.*, 2022; Wang and Sun, 2020]. Similar to the settings of those studies, we simulated these propagation processes until the state of 1% of users became active or infected in a network with 677,058 nodes and 828,546 edges[1].

5.2 Evaluation Metrics and Comparison Methods

By comprehensively comparing the proposed methods and the baseline methods, we use two evaluation metrics, i.e., the standard F1-score [Sokolova *et al.*, 2006] (F1) and average error distance (AED) [Dong *et al.*, 2022].

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{18}$$

$$\Delta_{\text{AED}} = \frac{1}{K}\sum_{k=1}^{K} d\left(r^*, r\right), \tag{19}$$

[1]Drawing from K' historical cascades $\mathcal{C}_k{=}(\mathcal{V}_k, \mathcal{E}_k, \mathcal{F}_k)$ in a social media platform, we construct the historical relationship network , which is a union graph by combining structural information of different cascades based on the same UIDs [Ramezani *et al.*, 2023].

Algorithm 2 Source inference process of GDFSL

Input: The optimized denoising module θ; An observed propagation cascade $\{\mathcal{G}^* = (\mathcal{V}, \mathcal{E}, \mathcal{F}, H^*)\}$.
Output: The predicted source $\hat{s}$.

$H^T \sim \mathcal{N}(\vec{\mathbf{1}}, \mathbf{I}), \hat{H}^T \leftarrow H^T$
$\mathcal{F}^* \leftarrow (\mathcal{F}, H^*)$
for $t = T, \ldots, 1$ **do**
  $z \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $z = 0$
  $\sigma^2_{t-1|t} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$ //Eq. (15)
  $\mu_{t-1|t} = \frac{1}{\sqrt{\alpha_t}}\Big(\hat{H}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{H}^t, t, \mathcal{V}, \mathcal{E}, \mathcal{F}^*)\Big) - \Big(\frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}(1-\sqrt{\bar{\alpha}_t})}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)} \frac{(1-\sqrt{\bar{\alpha}_{t-1}})(1-\sqrt{\alpha_t})(1-\sqrt{\bar{\alpha}_t})}{1-\bar{\alpha}_t}\Big)\vec{\mathbf{1}}$ //Eq. (17)
  $\hat{H}^{t-1} = \mu_{t-1|t} + \sigma_{t-1|t} \cdot z$
end for
return $\hat{s} = \text{argmax}_{v \in \mathcal{V}} \hat{H}^0(v)$

Statistic	Twitter15	Twitter16	Weibo
#users	480,987	289,675	2,856,741
#relationships	580,593	362,871	3,713,763
#cascades	1490	818	4664
#rumor cascades	370	205	2244
#non-rumor cascades	746	412	2082

Table 1: Statistics of the real-world datasets. Some invalid cascades do not belong to rumor or non-rumor cascades.

where *Recall* is the proportion of ground-truth sources that are successfully predicted. And *Precision* is the proportion of ground-truth sources in the predicted nodes. Here, *Recall* and *Precision* have the same weight in the standard F1. And $d(r^*, r)$ is the shortest distance between the predicted source and the ground-truth source.

In order to highlight the performance of the proposed methods, we choose eight SOTA methods for comparison. These SOTA methods includes DDMSL [Yan *et al.*, 2024], GIN-SD [Cheng *et al.*, 2024], Diff [Huang *et al.*, 2023], TGASI [Hou *et al.*, 2023], IVGD [Wang *et al.*, 2022], SL_VAE [Ling *et al.*, 2022], GCSSI [Dong *et al.*, 2022], and DRSDBFL [Wang *et al.*, 2024].

5.3 Performance on Synthetic Cascades

The source localization performance based on the synthetic simulation datasets is shown in Tab. 3. To ensure the generalizability and reusability of the proposed framework, we employ a convenient GCNs and GATs as the denoising module. For the performance on synthetic propagation datasets, when benchmarked against the optimal baseline DDMSL, GDFSL without utilizing the label information during training exhibits an average improvement of 7.4% in simulated datasets based on the F1 metric. Furthermore, using the average error distance as a metric, GDFSL also exhibits a smaller source error distance than all established baseline methods, reducing the error distance in the source evaluation by about 0.1 to 0.2 compared with the optimal baseline DDMSL. There are three key reasons for the significant improvement in real-world datasets: (1) The different signals of dynamic propagation features are learned and the influence of user profiles is dynamically enhanced in the source inference process. (2) The closed-form solutions of the forward and reversed diffusion process guarantees the interpretability of the proposed methods in source localization tasks. (3) User profiles, representing key social attributes, can positively guide the source localization process when explicitly embedded into the model architecture.

Dataset	Twitter15		Twitter16		Weibo	
Algorithm	F1	AED	F1	AED	F1	AED
DDMSL	<u>0.663</u>	<u>0.442</u>	<u>0.651</u>	<u>0.450</u>	<u>0.622</u>	<u>0.672</u>
GIN-SD	0.578	0.541	0.564	0.556	0.523	0.912
Diff	0.612	0.532	0.605	0.539	0.578	0.774
TGASI	0.515	0.637	0.511	0.642	0.497	0.921
IVGD	0.368	0.851	0.342	0.884	0.330	1.211
DRSDBFL	0.356	0.858	0.362	0.805	0.320	1.256
SL_VAE	0.355	0.855	0.342	0.871	0.346	1.271
GCSSI	0.228	1.031	0.224	1.062	0.255	1.413
GDFSL	**0.752**	**0.325**	**0.738**	**0.364**	**0.707**	**0.485**

Table 2: Source localization performance on the real-world cascades in Twitter15, Twitter16, Weibo. The bold values represent the best results, while underlined values denote the second-best.

5.4 Performance on Real-world Cascades

The source localization performance based on the real-world datasets is illustrated in Tab. 2. For traditional synthetic datasets, due to the inability to capture the dynamics of propagation data, GDFSL directly learns the random noise at each diffusion step. Nevertheless, GDFSL continues to exhibit superior performance. However, the performance improvements in real-world datasets are more pronounced than those in synthetic datasets. More specifically, the enhancements of GDFSL have increased from 7.4% of synthetic cascades to approximately 13.4% of real-world cascades. This increment in performance is due to the user-driven propagation scenarios in real-world networks, where the representation learning of user profiles provides a positive contribution to propagation source localization. While in the simulated propagation models, where user profiles do not exert the same unique influence as they do in the real world, the model is less capable of learning the relevance between complex dynamic and diverse profiles. Thus, the evaluation of practical characteristics by the attention module becomes less effective. This also underscores the necessity of source localization in real-world propagation scenarios.

5.5 Ablation Study

We further study the effectiveness of the components designed within our proposed methods to validate their contributions to source detection performance. Here we present the ablation study of GDFSL in the Twitter15 dataset. The critical ablation settings include:

Dynamic	HDP		TDP		RDP		LT		IC		SI	
Algorithm	F1	AED	F1	AED	F1	AED	F1	AED	F1	AED	F1	AED
DDMSL	0.568	0.720	<u>0.685</u>	**0.630**	0.698	0.710	<u>0.574</u>	<u>0.760</u>	<u>0.602</u>	<u>0.620</u>	<u>0.591</u>	<u>0.600</u>
GIN-SD	<u>0.582</u>	<u>0.690</u>	0.611	0.680	0.645	0.760	0.554	0.800	0.517	0.750	0.506	0.690
Diff	0.578	0.710	0.669	0.660	<u>0.702</u>	<u>0.710</u>	0.565	0.770	0.572	0.670	0.553	0.660
TGASI	0.556	0.760	0.584	0.790	0.610	0.780	0.528	0.830	0.532	0.800	0.544	0.770
IVGD	0.413	0.950	0.387	1.110	0.405	1.260	0.367	1.200	0.357	1.250	0.346	1.300
DRSDBFL	0.432	0.920	0.358	1.150	0.403	1.280	0.350	1.220	0.366	1.250	0.361	1.220
SL_VAE	0.314	1.060	0.333	1.200	0.327	1.350	0.351	1.220	0.388	1.220	0.355	1.250
GCSSI	0.207	1.360	0.215	1.350	0.188	1.520	0.105	1.550	0.146	1.500	0.122	1.550
GDFSL	**0.634**	**0.590**	**0.688**	<u>0.650</u>	**0.715**	**0.650**	**0.585**	**0.750**	**0.648**	**0.590**	**0.635**	**0.560**

Table 3: Source localization performance based on six different propagation dynamics. The bold values represent the best results, while underlined values denote the second-best.

- "$\mathcal{N}(\vec{\mathbf{1}}, \mathbf{I}) \rightarrow N(0, \mathbf{I})$" uses the standard Gaussian noise in the DDPM instead of the non-zero mean at the T step.
- "$-\nabla_\theta \left\| H^* - \hat{H}^* \right\|^2$" does not use the observation constraints (i.e., lines **12-19** in Alg. 1) as a training loss item.
- "$-\nabla_\theta \left\| \epsilon - \epsilon_\theta \right\|^2$" does not use the noise estimation loss (i.e., lines **2-11** in Alg. 1) as a training loss item.
- "-Att" removes the graph attention module, and only the GCNs are available.
- "-PE" does not consider the time step index in the denoising module.

As shown in Tab. 4, it will lead to a performance decrease or a delay in the convergence training speed no matter removing or exchanging any critical modules. As for replacing the non-zero mean with a standard Gaussian noise ($\mathcal{N}(\vec{\mathbf{1}}, \mathbf{I}) \rightarrow N(0, \mathbf{I})$), the non-zero mean is crucial as it aligns the diffusion process with the fully infected final state, which is essential for accurate source detection. Otherwise, the interpretability of the diffusion process will be lost. As for excluding the observation constraints $\nabla_\theta \left\| H^* - \hat{H}^* \right\|^2$, the observation constraints ensure that the generated snapshots are consistent with the observed data, which is vital for the GDFSL model to learn accurate representations. As for excluding the noise estimation loss $\nabla_\theta \left\| \epsilon - \epsilon_\theta \right\|^2$ from the training loss, estimating the noise accurately is fundamental to the denoising process, and removing it results in an incomplete diffusion training process.

Furthermore, the influence of the diffusion framework is found to be greater than that of the denoising module. This highlights the crucial role of the proposed diffusion framework in ensuring effective source localization. The results indicate that the GDFSL framework benefits most from its novelly designed forward and reverse processes.

Modules	*Variants*	F1	AED	*Epoch*
Diffusion	$\mathcal{N}(\vec{\mathbf{1}}, \mathbf{I}) \rightarrow N(0, \mathbf{I})$	0.417	0.872	5
	$-\nabla_\theta \left\| H^* - \hat{H}^* \right\|^2$	0.692	0.408	5
	$-\nabla_\theta \left\| \epsilon - \epsilon_\theta \right\|^2$	0.676	0.437	4
Denoising	-Att	0.713	0.396	4
	-PE	0.725	0.364	**3**
Origin	GDFSL	**0.752**	**0.325**	**3**

Table 4: The performance evaluation of variant models from GDFSL in the Twitter15 dataset. - signifies that the corresponding modules are removed or masked with zeros. And the symbol $\rightarrow$ indicates that the original module is replaced by a new part.

6 Conclusion

In this paper, we propose the GDFSL, a flexible and generalized framework to source localization tasks in diverse propagation scenarios. GDFSL overcomes the limitations of traditional methods that rely on specific propagation models by leveraging noise to learn the unique propagation pattern of each dynamic. Through the redefined forward process and an interpretable closed-form reverse diffusion solution, GDFSL accurately infers the source from observed data without relying on adequate labeled historical datasets for training. Experimental results demonstrate that GDFSL outperforms existing SOTA localization methods across various synthetic and real-world scenarios, highlighting its effectiveness and transferability.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 62271411, U22A2098, 62471403, 62261136549), the Technological Innovation Team of Shaanxi Province (No. 2025RS-CXTD-009), the International Cooperation Project of Shaanxi Province (No. 2025GH-YBXM-017), the Fundamental Research Funds for the Central Universities (Nos. G2024WD0151, D5000240309).

References

[Cheng *et al.*, 2024] Le Cheng, Peican Zhu, Keke Tang, Chao Gao, and Zhen Wang. Gin-sd: source detection in graphs with incomplete nodes via positional encoding and attentive fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 55–63, 2024.

[Dong *et al.*, 2019] Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and Guohui Li. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 569–578, 2019.

[Dong *et al.*, 2022] Ming Dong, Bolong Zheng, Guohui Li, Chenliang Li, Kai Zheng, and Xiaofang Zhou. Wavefront-based multiple rumor sources identification by multi-task learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1068–1078, 2022.

[Ellison, 2024] Glenn Ellison. Implications of heterogeneous sir models for analyses of covid-19. *Review of Economic Design*, pages 1–37, 2024.

[Goldenberg *et al.*, 2001] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[Granovetter, 1978] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hou *et al.*, 2023] Dongpeng Hou, Zhen Wang, Chao Gao, and Xuelong Li. Sequential attention source identification based on feature representation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4794–4802, 2023.

[Hou *et al.*, 2024a] Dongpeng Hou, Chao Gao, Zhen Wang, Xiaoyu Li, and Xuelong Li. Random full-order-coverage based rapid source localization with limited observations for large-scale networks. *IEEE Transactions on Network Science and Engineering*, 2024.

[Hou *et al.*, 2024b] Dongpeng Hou, Shu Yin, Chao Gao, Xianghua Li, and Zhen Wang. Propagation dynamics of rumor vs. non-rumor across multiple social media platforms driven by user characteristics. *arXiv preprint arXiv:2401.17840*, 2024.

[Hou *et al.*, 2025] Dongpeng Hou, Chao Gao, Zhen Wang, and Xuelong Li. Fgssi: A feature-enhanced framework with transferability for sequential source identification. *IEEE Transactions on Dependable and Secure Computing*, 2025.

[Huang *et al.*, 2018] Chunlin Huang, Xingwu Liu, Minghua Deng, Yang Zhou, and Dongbo Bu. A survey on algorithms for epidemic source identification on complex networks. *Chinese Journal of Computers*, 41(06):1156–1179, 2018.

[Huang *et al.*, 2023] Bosong Huang, Weihao Yu, Ruzhong Xie, Jing Xiao, and Jin Huang. Two-stage denoising diffusion model for source localization in graph inverse problems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 325–340, 2023.

[Jiang *et al.*, 2016] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, 19(1):465–481, 2016.

[Karrer and Newman, 2010] Brian Karrer and Mark EJ Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101, 2010.

[Ling *et al.*, 2022] Chen Ling, Junji Jiang, Junxiang Wang, and Zhao Liang. 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*, pages 1010–1020, 2022.

[Liu *et al.*, 2015] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870, 2015.

[Ma *et al.*, 2016] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Cha Meeyoung. Detecting rumors from microblogs with recurrent neural networks. In *The 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824, 2016.

[Ma *et al.*, 2017] Jing Ma, Wei Gao, and Kam-Fai Wong. 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, pages 708–717, 2017.

[Paluch *et al.*, 2021] Robert Paluch, Łukasz Gajewski, Krzysztof Suchecki, and Bolesław Szymański. Enhancing maximum likelihood estimation of infection source localization. In *Simplicity of Complexity in Economic and Social Systems*, pages 21–41, 2021.

[Ramezani *et al.*, 2023] Maryam Ramezani, Aryan Ahadinia, Amirmohammad Ziaei Bideh, and Hamid R Rabiee. Joint inference of diffusion and structure in partially observed social networks using coupled matrix factorization. *ACM Transactions on Knowledge Discovery from Data*, 17(9):1–28, 2023.

[Shah and Zaman, 2010] Devavrat Shah and Tauhid Zaman. Detecting sources of computer viruses in networks: theory and experiment. In *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 203–214, 2010.

[Shah and Zaman, 2011] Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011.

[Shu *et al.*, 2021] Xincheng Shu, Bin Yu, Zhongyuan Ruan, Qingpeng Zhang, and Qi Xuan. Information source estimation with multi-channel graph neural network. *Graph Data Mining: Algorithm, Security and Application*, pages 1–27, 2021.

[Sokolova *et al.*, 2006] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021, 2006.

[Tang *et al.*, 2018] Wenchang Tang, Feng Ji, and Wee Peng Tay. Estimating infection sources in networks using partial timestamps. *IEEE Transactions on Information Forensics and Security*, 13(12):3035–3049, 2018.

[Wang and Sun, 2020] Hongjue Wang and Kaijia Sun. Locating source of heterogeneous propagation model by universal algorithm. *Europhysics Letters*, 131(4):48001, 2020.

[Wang *et al.*, 2017] Zheng Wang, Chaokun Wang, Jisheng Pei, and Xiaojun Ye. Multiple source detection without knowing the underlying propagation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 217–223, 2017.

[Wang *et al.*, 2022] Junxiang Wang, Junji Jiang, and Liang Zhao. An invertible graph diffusion neural network for source localization. In *Proceedings of the ACM Web Conference*, pages 1058–1069, 2022.

[Wang *et al.*, 2024] Ranran Wang, Yin Zhang, Wenchao Wan, Min Chen, and Mohsen Guizani. Distributed rumor source detection via boosted federated learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5986–6001, 2024.

[Yan *et al.*, 2024] Xin Yan, Hui Fang, and Qiang He. Diffusion model for graph inverse problems: Towards effective source localization on complex networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yang *et al.*, 2020] Fan Yang, Shuhong Yang, Yong Peng, Yabing Yao, Zhiwen Wang, Houjun Li, Jingxian Liu, Ruisheng Zhang, and Chungui Li. Locating the propagation source in complex networks with a direction-induced search based gaussian estimator. *Knowledge-Based Systems*, 195:105674, 2020.

[Zang *et al.*, 2015] Wenyu Zang, Chuan Zhou, Li Guo, and Peng Zhang. Topic-aware source locating in social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 141–142, 2015.

[Zhu and Ying, 2014] Kai Zhu and Lei Ying. Information source detection in the sir model: A sample-path-based approach. *IEEE/ACM Transactions on Networking*, 24(1):408–421, 2014.

[Zhu *et al.*, 2017] Kai Zhu, Zhen Chen, and Lei Ying. Catch'em all: Locating multiple diffusion sources in networks with partial observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1676–1682, 2017.

[Zhu *et al.*, 2022] Junyou Zhu, Chunyu Wang, Chao Gao, Fan Zhang, Zhen Wang, and Xuelong Li. Community detection in graph: An embedding method. *IEEE Transactions on Network Science and Engineering*, 9(2):689–702, 2022.

[Zhu *et al.*, 2024] Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Jürgen Kurths. Propagation structure-aware graph transformer for robust and interpretable fake news detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4652–4663, 2024.